Psychometrika (2025), **0**, 1–30
doi:10.1007/psy.2024.27

PSYCHOMETRIC SOCIETY

APPLICATION AND CASE STUDIES

Comparing Functional Trend and Learning among Groups in Intensive Binary Longitudinal Eye-Tracking Data using By-Variable Smooth Functions of GAMM

Sun-Joo Cho[1], Sarah Brown-Schmidt[1], Sharice Clough[2,3] and Melissa C. Duff[2]

[1]Vanderbilt University; [2]Vanderbilt University Medical Center; [3]The Max Planck Institute for Psycholinguistics

Corresponding author: Sun-Joo Cho; Email: sj.cho@vanderbilt.edu

(Received 7 November 2023; accepted 5 June 2024)

This manuscript is part of the special section Model Identification and Estimation for Longitudinal Data in Practice. We thank Drs. Carolyn J. Anderson and Donald Hedeker for serving as co-Guest Editors.

Abstract

This paper presents a model specification for group comparisons regarding a functional trend over time within a trial and learning across a series of trials in intensive binary longitudinal eye-tracking data. The functional trend and learning effects are modeled using by-variable smooth functions. This model specification is formulated as a generalized additive mixed model, which allowed for the use of the freely available mgcv package (Wood in Package 'mgcv.' https://cran.r-project.org/web/packages/mgcv/mgcv.pdf, 2023) in R. The model specification was applied to intensive binary longitudinal eye-tracking data, where the questions of interest concern differences between individuals with and without brain injury in their real-time language comprehension and how this affects their learning over time. The results of the simulation study show that the model parameters are recovered well and the by-variable smooth functions are adequately predicted in the same condition as those found in the application.

Keywords: by-variable smooth function; eye-tracking data; generalized additive mixed model; group comparisons; intensive binary longitudinal data

1. Introduction

1.1. Empirical Motivation

This paper is motivated by the need for statistical analysis methods that can facilitate group comparisons regarding trends in eye-fixation data over time and learning across a series of trials in intensive (many time points; e.g., 252 time points) binary longitudinal eye-tracking data. The present work uses an established paradigm in which eye fixations provide insight into the language processing, popularly known as the *visual world paradigm* (Tanenhaus et al., 1995). Visual world paradigm involves tracking eye gaze to images in visual displays as research participants produce or interpret language that is related to the viewed images. The time course with which participants gaze at, e.g., a picture of a "sandwich" as they interpret the sentence "The girl will eat the sandwich", as opposed to an "apple" or a "piano" can offer insights into the cognitive mechanisms involved in understanding language as it unfolds over time. A primary goal of the present analysis is to leverage fixation data from the visual world paradigm

to examine differences between individuals with and without brain injury in their ability to understand language in the moment, and how this supports learning over time.

1.2. Data Complexities and Limitations of Existing Model Specification

In comparing trend and learning effects among groups, it is crucial to acknowledge the following data complexities in intensive binary longitudinal eye-tracking data. First, there exists a *temporal dependency* in such data, which reflects the stability of fixation on a given fixation area over time. Cho et al. (2018) addressed this issue by considering the observation-driven autoregressive (AR) effect (Cox & Snell, 1989, pp. 100–101) for which a target fixation (outcome variable) at the current time point is determined by target fixations at previous time points. Cho et al. (2018) demonstrated via a simulation study that neglecting this AR effect in analyses led to biased estimates and standard errors of focal parameters (e.g., experimental condition effects) in intensive binary longitudinal eye-tracking data. Second, a common research question concerns the fixation *trend* over time in intensive binary longitudinal eye-tracking data. A time slope parameter can be incorporated to capture the presence and steepness of the fixation trend, which can indicate slower or faster identification of a target area in question. Third, there are different uses of the *trial* variable in experimental designs. The trial variable can either serve as a unique identifier for each trial or represent an ordered trial variable representing the sequence in which each participant completed a series of trials. When the trial variable is used to identify unique trials, a random intercept is considered to account for clustering due to trials. However, when comparing learning effects across groups, the trial variable should represent the ordered trial number.

In practice, growth curve analysis has been used with proportion-based measures of gaze[1] to model patterns in the proportion of gaze to a given interest area over time using linear, quadratic, and/or cubic polynomial terms in the growth curve model (e.g., Baker & Love, 2021; Brown et al., 2011; Hadar et al., 2016). In addition, previous research has addressed the modeling of trend over time in intensive binary longitudinal eye-tracking data. For instance, Cho et al. (2020) accounted for time slope parameters of linear and quadratic polynomials of time covariates within a generalized linear mixed-effects model (GLMM). Furthermore, target fixations over time were fit by a four-parameter logistic function that captures the asymptotes, slope, and crossover point (Ito and Knoeferle, 2022; Oleson et al., 2017). Using polynomials and nonlinear functions offers the potential advantage of assigning a meaningful interpretation to each parameter within the study's context (Ram and Grimm, 2007). For example, the linear term and the quadratic term in the polynomials reflect a monotonic change in gaze and the symmetric rise and fall rate around a central inflection point, respectively (e.g., Mirman et al., 2008).

As another approach, a *functional* perspective (Ramsay and Silverman, 2002, 2005) can be adopted to capture the intrinsic structure of the data, focusing on its underlying nature rather than its explicit form such as the polynomials and nonlinear functions. For the current study, we chose a functional perspective using a generalized additive mixed model (GAMM; Lin & Zhang, 1999; Wood, 2017) because it does not require a priori knowledge of the functional form for modeling group-specific trends over time and learning across multiple trials. In psycholinguistics, GAMMs have been widely applied to investigate the temporal dynamics of continuous time series data and group differences using by-variable smooth functions (Baayen et al., 2017, 2018, 2022; Chuang et al., 2021; Heitmeier et al., 2023; Ito & Knoeferle, 2022; Porretta et al., 2017; van Rij et al., 2019; Wieling, 2018). As an example of these studies, Baayen et al. (2017) demonstrated the use of a smooth function of trial orders to model the functional trend effect while modeling AR in the residuals, and the use of by-variable smooth functions to investigate gender-group differences in word frequency in continuous (log-transformed) response time data within a GAMM framework. In addition, Baayen et al. (2022) presented a GAMM to facilitate individual-specific functional trend effects using by-variable smooth functions, in detecting the fixed

[1] As an example, Akhavan et al. (2023) defined the proportion by dividing the total duration of gazes on a particular area of interest by the total duration of a trial.

https://doi.org/10.1007/psy.2024.27 Published online by Cambridge University Press

effects of within- and between-subjects factors. In psychometrics, Sørensen et al. (2023) presented generalized additive latent and mixed models (GALAMMs) for longitudinal data with outcome and latent variables depending smoothly on observed variables. In a recent study, Cho et al. (2022) utilized a smooth function of time to model the *mean* functional trend and trial order effects across persons and items in intensive binary longitudinal eye-tracking data within the GAMM. However, prior studies employing GAMMs have not investigated the use of by-variable smooth functions for time and trial orders in group comparisons (e.g., between persons with and without brain injury in an experimental condition having a within-subjects factor) for binary longitudinal data. This gap highlights the need for further methodological development, including modeling AR effects and their variability across persons and items and probing interactions between groups and the experimental condition based on results of the by-variable smooth function within GAMMs.

Functional data analysis (FDA), as introduced by Ramsay and Silverman (2002, 2005), shares a strong connection with GAMM when it comes to modeling functional effects (see examples for relations between FDA and GAMM in Wood 2017, pp. 390–397). FDA and its extensions to mixed-effects modeling (known as functional mixed-effects modeling [FMEM], Guo 2002) have been applied to longitudinal or time series data.[2] For example, Fine et al. (2019) presented the application of the FMEM to analyze continuous longitudinal data, enabling the consideration of the population mean and individual differences in trends over time. In addition, Durbán et al. (2005) and Suk et al. (2019) applied the mixed-effects penalized spline (as a special case of FMEM or GAMM) to model group-specific (e.g., sex) and individual-level functional trend over time in the continuous longitudinal data. As another example, Staicu et al. (2020) presented the longitudinal dynamic functional regression (as a special case of FMEM) for both continuous and binary longitudinal data to model group-specific (e.g., young vs. old sows) trend and time-varying covariate effects using smooth functions. However, these studies did not consider AR effects and multivariate binary longitudinal data (e.g., from multiple trials and items), which are commonly found in visual world eye-tracking data. In this study, GAMM was chosen over FDA or FMEM because the by-variable smooth functions to model group-specific functional effects of interest were developed within a GAMM framework.

1.3. Study Purposes and Novel Methodological Contributions

The current study has three main purposes. First, we aim to present a model specification for comparing functional trends over time and learning over multiple trials among groups, along with the AR effect, in intensive binary longitudinal eye-tracking data. This is achieved by utilizing by-variable smooth functions of GAMM. The by-variable smooth function estimates a smooth function for each diagnostic group, such as non-injured comparison (NC) participants or participants with traumatic brain injury (TBI) in our motivating data set, in an experimental condition. To the best of our knowledge, these by-variable smooth functions have not been applied to intensive *binary* longitudinal data for the purpose of group comparisons. In addition, psycholinguistics research has widely advocated for the use of crossed random person and item effects to simultaneously account for person and item heterogeneity (e.g., Baayen et al., 2008). Thus, such heterogeneity is considered in the model specification. Furthermore, we demonstrate how to model variability in trend and learning and in AR effects using random slopes within GAMM, which may differ across individuals and/or items.

Second, when detecting group differences in trend and learning in the experiment, it is important to probe interactions between diagnostic group (TBI vs. NC) and conditions, and to identify the time and trial order ranges in which differences in the effects on target fixations can be observed between the diagnostic groups. Therefore, we present a procedure to detect the interaction and the ranges for the

[2]Both longitudinal data and time series data involve observations taken over time. However, in the literature, they are often used differently in terms of focus, purpose, and structure in analyses. For the eye-tracking data we focus on in the current study, we use the terms 'intensive (many time points) binary longitudinal data' and 'binary time-series data' interchangeably.

https://doi.org/10.1007/psy.2024.27 Published online by Cambridge University Press

differences in this study, based on the results of the by-variable functions. We employ statistical testing methods and present differences-between-smooths plots as part of this purpose.

For parameter estimation, the bam function of the mgcv package (Wood, 2023) in R (R Core Team, 2023) is used. While the bam function is designed to handle a wide range of GAMM specifications, it may pose challenges for researchers aiming to use it for the specified model from our first aim. Hence, our third purpose is to provide a detailed explanation of how parameter estimation is implemented using the mgcv package.

The specified model is applied to a motivating example data set, hypothesizing that there are different profiles in trend and learning between the diagnostic group (TBI vs. NC) in the data set. In the motivating example data set, different options for how to code "trials" in the experiment are discussed. Additionally, a simulation study is conducted to evaluate the recovery of parameters of the specified model. These aspects have not been previously demonstrated in the literature.

This paper is organized as follows. In Sect. 2, the empirical study that motivated the current paper is described. In Sect. 3, a model is specified to incorporate a by-variable smooth function of time and trial orders to model functional trends over time and learning across a series of trials among groups. In Sect. 4, the specified model is illustrated using an empirical data set. In Sect. 5, the simulation study is presented to evaluate parameter recovery, and type I error rate and power of testing for a by-variable smooth function. In Sect. 6, we end with a summary and a discussion.

2. Motivating Empirical Study

In this section, the motivating empirical study is explained. In addition, the different coding options for trials within the experiment are discussed.

2.1. Samples and Experimental Design

The dataset (Clough et al., 2023) is from a study of 45 individuals with moderate–severe TBI and 44 NC participants who were matched to the individuals in the TBI group on age, sex, and education (the data from one person in the NC group was lost due to equipment failure). TBI causes heterogeneous deficits in cognition, including language deficits (Covington and Duff, 2021; Dahlberg et al., 2006). The study used a variant of the visual world paradigm (Tanenhaus et al., 1995), in which eye-tracked participants completed a series of trials where they clicked on objects that were mentioned in a sentence. On each trial, participants viewed 4 images, such as a "sandwich", an "apple", a "piano", and a "guitar." A video in the center of the screen featured a person who said the corresponding sentence, e.g., "The girl will eat the very nice sandwich", which always contained a verb that was potentially consistent with two of the objects on-screen, the named target (sandwich) and a competitor (apple)" (see Fig. 1).

Critically, on half the trials, the speaker used a *representative gesture* that was consistent with the target (e.g., a sandwich holding gesture); this gesture was produced at onset of the verb phrase (e.g., will eat) before the spoken target word "sandwich" and thus provided an early cue to speaker meaning. On the other half of trials, the speaker used a non-informative *grooming movement* (e.g., scratching their arm). The eye-tracking data are analyzed beginning 180 ms after the onset of the gesture or grooming movement "stroke" in the video, allowing us to capture eye gaze made in response to the combination of verb and gesture/grooming movement. The end of the analysis window was the average onset of the spoken target word (e.g., "sandwich"), which was 2700 ms later. The analysis window is delayed by 200 ms due to the time needed to launch an eye movement in response to an external stimulus, minus 20 ms needed in order to calculate a baseline for the AR term in the model (Cho et al., 2018).

The eye-tracking data are coded in binary form, reflecting whether or not the participant was fixating the target (e.g., "sandwich") at each 10 ms time point in time during the analysis window. Each participant completed 240 trials where they viewed a scene with 4 pictures, heard an associated sentence produced by a speaker in a video, and were tasked with clicking on the associated picture (e.g., the "sandwich"). As participants completed the 240 trials, the order of those trials (from 1 to 240) may

https://doi.org/10.1007/psy.2024.27 Published online by Cambridge University Press

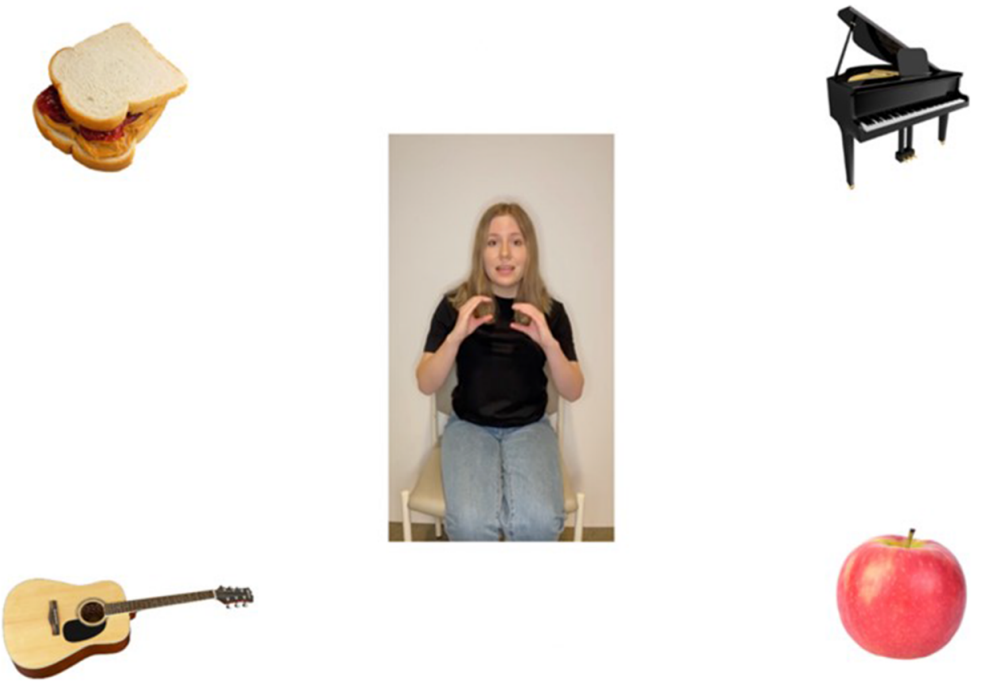

Figure 1. Empirical study: Example trial including sandwich (target), apple, piano, and guitar images, and a video still of a person in the gesture condition making a sandwich gesture.

be relevant to learning within the experiment about the task, and thus we may expect performance to improve over the course of the 240 trials (or alternatively, if fatigue sets in, for performance to decline across the 240 trials). When we refer to *trial order*, we mean the order of the 240 trials that a given participant completed. On each trial, the sentence referred to one of four pictures, termed the *target* picture (e.g., the "sandwich"). Critically, the scenes always contained a *competitor* picture (e.g., the "apple") which shared some affordances with the target such that both were potential direct objects of the verb in the sentence. For example, both the target ("sandwich") and competitor ("apple") are possible direct objects of the verb "eat." Across trials, we varied whether, for a given pair of objects which was the target and which was the competitor, e.g., on some trials the "sandwich" was the target (and the "apple" was the competitor) and on other trials the "apple" was the target (and the "sandwich" was the competitor). There was a total of 40 pairs of target pictures that shared a verb; we refer to these as the 40 different *items* in the analysis, and given inter-item differences in features like the imageability or frequency of the verb and the degree to which it evoked the associated objects, we can expect dependency in item responses due to item clustering. For each of the 40 items, a version of the sentence was recorded in both the gesture and grooming movement conditions and for both possible nouns (i.e., 160 unique sentence-hand movement combinations). We recorded four speakers (two male, two female) producing each of these 160 sentences, resulting in a total of 640 unique possible *trials*. These 640 trials were put into a single randomized order and divided into eight blocks of 80 trials each. From those 8 blocks, a Latin square sampling design was used to create eight stimulus lists of 240 trials. Participants were randomly assigned to one of these stimulus lists. The trial number is a unique identifier that captures which of 640 specific combinations of the 80 items (e.g., sandwich, apple, etc.), gesture condition (with vs. without gesture), and speaker identity were featured in the video on a given trial. Note that because there were only 80 unique items (e.g., "sandwich/apple; piano/guitar", etc.) but each participant completed a sequence of 240 trials, each person experienced each item multiple times.

https://doi.org/10.1007/psy.2024.27 Published online by Cambridge University Press

In sum, the empirical data set has a nested and cross-classified multilevel structure: the 252 time points at level 1 are nested within 240 trial orders at level 2, which are cross-classified by 89 persons (44 NC participants and 45 TBI participants) and 40 items at level 3. The resulting data have a total of 5,382,720 binary data points ($252 \times 240 \times 89 = 5{,}382{,}720$).

In what follows, we use the term *condition* to refer to the distinction between gesture and grooming movements in experiment. We use *diagnostic group* to describe the distinction between NC and TBI. Finally, we use the term *group* for the categorical covariate that forms a by-variable smooth function. For instance, such groups include NC participants in a grooming movement condition, NC participants in a gesture condition, TBI participants in a grooming movement condition, and TBI participants in a gesture condition.

2.2. Empirical Research Questions and Hypotheses

The substantive research question of interest is how moderate–severe TBI impacts online language processing in rich contexts, and if and how disruptions in online language processing scale up over time to impact learning over longer time scales. Based on prior findings using similar methods (Altmann and Kamide, 1999; Cho et al., 2020), we expect to observe a significant positive trend effect such that over 252 time points within a trial, the probability of a target fixation (coded as 1) compared to a non-target fixation (coded as 0) increases across time points. This effect reflects the successful interpretation of the sentence over time; as the person interprets the sentence, they become more likely to fixate the target referent (e.g., "sandwich" in Fig. 1). While we expect target fixations to increase across time, nonlinearity in the trend can be expected as once the target has been identified on most trials, there is likely to be an asymptotic effect, and in some cases fixations to the target may drop after persons locate and click the target and then look away (see Yoon & Brown-Schmidt, 2018). Critically, success in this task requires integrating the unfolding linguistic signal with the speaker's gesture, and with the associated visual scene in order to direct fixations to the target referent. Impairments in the ability to process language, gesture, the visual scene, or the integration of any of these elements in TBI would likely delay target identification, a result which we speculate would be reflected in a shallower rise in target fixations over time (a shallower trend effect in TBI).

We also test for a learning effect across the 240 trial orders, which we hypothesize to be reflected in an increased probability of a target fixation as trial order increases. Such an effect would represent learning how to more efficiently process speech and gesture, and locate the target image within the scene as experience with the task increases across trials. An asymptotic effect in learning may reflect the fact that at some point, individuals maximize their ability to understand and quickly process the speech and gesture. If TBI impairs the ability to learn within this task, we would expect a slower increase in target fixations across the 240 trial orders of the experiment, or even a decrease if participants with TBI experience more fatigue as the task progresses, compared to non-injured participants.

Lastly, an interaction between trend and learning effects can be expected if efficient processing of speech and gesture at the trial level (trend) supports efficient learning across trials. If so, we would expect that the participant group (or individual participants) that shows stronger trend effects (e.g., larger increases in target fixations over time within a trial) to show more learning across trials. One potential outcome is that online processing is impaired in TBI (reflected in shallower trend effects within trials compared to non-injured participants), and that this in turn is associated with weaker learning across trials. Another potential outcome is that online processing is intact in TBI (reflected in equivalent trend effects within trials compared to non-injured participants), but that TBI impairs the ability to translate in-the-moment processing into longer-term learning gains. If so, participants with TBI may show equivalent trend effects as non-injured comparison participants, but nonetheless show weaker learning effects across trials. However, we do not have a strong hypothesis regarding the form of the group-specific trend and learning effects. We assume that a smooth function could approximate the underlying nature of the group-specific trend and learning effectively.

https://doi.org/10.1007/psy.2024.27 Published online by Cambridge University Press

3. Methods

In this section, GAMM specification is provided to answer the substantive research question. Subsequently, parameter estimation, model comparisons and evaluation, and testing smooth functions and plotting differences-between-smooths are described.

3.1. Model Specification

Denote a binary observation at an equally spaced time point t ($t = 1,\ldots,T$) from trial order l ($l = 1,\ldots,L$), person j ($j = 1,\ldots,J$), and item i ($i = 1,\ldots,I$) in group g ($g = 1,\ldots,G$) by y_{tljig}. The first-order AR (AR1) is considered in model specification for illustrative purpose (also found in the empirical study). Below, three models are specified to compare functional trend and learning effects among groups. (a) Model A: fixed AR1 and group-specific functional trend and learning effects using by-variable smooth functions, (b) Model B: adding variability in AR across participants and items using random slopes to Model A, and (c) Model C: adding variability in trend and learning effects across persons and items using random slopes to Model B. Focal effects of interest are group-specific functional trend and learning effects, controlling for variability in AR, trend, and learning effects across persons and items.

Model A

The model with fixed AR1 and group-level functional trend and learning effects is written as

$$\begin{aligned}\text{logit}[P(y_{tljig}=1|y_{(t-1)ljig},\mathbf{X}_j,time_t,trialorder_l,\alpha_1,\boldsymbol{\alpha},\zeta_1,\boldsymbol{\zeta},\theta_j,\beta_i)] \\ =\alpha_1+\boldsymbol{\alpha}\mathbf{X}_j+\zeta_1 y_{(t-1)ljig}+\boldsymbol{\zeta}(\mathbf{X}_j y_{(t-1)ljig})+\theta_j+\beta_i \\ +f(time_t)(group_j=g)+f(trialorder_l)(group_j=g) \\ +f(time_t,trialorder_l)(group_j=g),\end{aligned} \tag{1}$$

where y_{tljig} is a binary response; $y_{tljig} = 1$ when person j looks at a target and $y_{tlji} = 0$ otherwise, $\mathbf{X}_j$ is a vector of the dummy-coded group covariate, $y_{(t-1)ljig}$ is an AR1 covariate (the first order of the lagged response variable), $group_j$ is a categorical group covariate ($group_j = 1,\ldots,G$), $time_t$ is a time covariate, $trialorder_l$ is a trial order covariate, α_1 is a fixed intercept for a reference group, $\boldsymbol{\alpha} = [\alpha_2,\ldots,\alpha_G]'$ is a vector of the difference between a reference group and another group, ζ_1 is a fixed AR1 effect for a reference group; the AR1 effect can be interpreted as the log-odds ratio for the current response due to the previous response changing from 0 to 1, $\boldsymbol{\zeta} = [\zeta_2,\ldots,\zeta_G]'$ is a vector of fixed AR1 effects for the difference between a reference group and another group, θ_j is a random person intercept to allow for individual differences, β_i is a random item intercept to allow for differences between the items, $f(time_t)(group_j = g)$ is a by-variable smooth function of $time_t$ to model a group-specific functional trend effect, $f(trialorder_l)(group_j = g)$ is a by-variable smooth function of $trialorder_l$ to model a group-specific functional learning effect, and $f(time_t, trialorder_l)(group_j = g)$ is a two-dimensional by-variable smooth function of $(time_t, trialorder_l)$ to model group-specific functional interaction between trend and learning effect (as in the smooth analysis of variance [ANOVA] model (e.g., Gu, 2013)). Normality is assumed for θ_j and β_i, respectively: $\theta_j \sim N(0,\sigma_\theta)$ and $\beta_i \sim N(0,\sigma_\beta)$. In the presence of both trend and AR effects in the model, our interpretation is that the AR1 effect primarily captures the short-term dependencies in the data (around the trend), while the trend represents the long-term direction or pattern.

In Eq. 1, a by-variable smooth function ($f(time_t)(group_j = g)$ or $f(trialorder_l)(group_j = g)$) is specified as a weighted sum of a set of basis functions over the covariate ($time_t$ or $trialorder_l$) for each group. For $f(time_t)(group_j = g)$ as an example,

$$f(time_t)(group_j=g)=\sum_{k=1}^{K_1}\delta_{k.g}b_{k.g}(time_t)(group_j=g), \tag{2}$$

https://doi.org/10.1007/psy.2024.27 Published online by Cambridge University Press

where k is an index for a kth basis function ($k = 1,\ldots,K_1$), $\delta_{k.g}$ is a basis coefficient for group g, and $b_{k.g}(time_t)(group_j = g)$ is the kth basis function for group g. The by-variable smooth function is identified with the constraint that the function sum over the observed time values across observations is 0.

In addition, a by-variable tensor smooth (e.g., Wood, 2017) is used for a two-dimensional by-variable smooth function ($f(time_t, trialorder_l)(group_j = g)$) in Eq. 1 because it has a property that a unit change in one variable is equivalent to a unit change in another variable. The by-variable tensor smooth function can be constructed as a weighted sum of a set of basis functions defined over the covariates of $time_t$ and $trialorder_l$ for group g:

$$f(time_t, trialorder_l)(group_j = g) = \sum_{k}^{K_2}\sum_{k'}^{K_2'} \gamma_{lkk'} b_{lkk'}(time_t, trialorder_l)(group_j = g), \tag{3}$$

$$= \sum_{k}^{K_2}\sum_{k'}^{K_2'} \gamma_{lkk'} b_{lk}(time_t) b_{lk'}(trialorder_l)(group_j = g), \tag{4}$$

where k is an index for a basis function ($k = 1,\ldots,K_2$) for a time covariate $time_t$, k' is an index for a basis function ($k' = 1,\ldots,K_2'$) for a trial order covariate $trialorder_l$, $\gamma_{lkk'}$ is a basis coefficient, and $b_{lkk'}(time_t, trialorder_l)$ is a bivariate basis function. For the two-dimensional by-variable smooth function, the bivariate basis function is $b_{lkk'}(time_t, trialorder_l) = b_{lk}(time_t) b_{lk'}(trialorder_l)$, which is a tensor product of univariate basis functions in $time_t$ and $trialorder_l$ directions. For identification, the marginal smooths of a tensor product are created with sum-to-zero identifiability constraints prior to constructing the tensor product basis.

Model B: Adding Varying AR1 Effects across Persons and Items to Model A

The random slopes of $y_{(t-1)ljig}$ (ζ_{1j} and ζ_{2i}) are added to Model A to model variability in AR1 the effect across persons and items, respectively:

$$\begin{aligned}
&\text{logit}[P(y_{tljig} = 1 | y_{(t-1)ljig}, \mathbf{X}_j, time_t, trialorder_l, \alpha_1, \boldsymbol{\alpha}, \zeta_1, \boldsymbol{\zeta}, \zeta_{1j}, \zeta_{2i}, \theta_j, \beta_i)] \\
&= \alpha_1 + \boldsymbol{\alpha}\mathbf{X}_j + \zeta_1 y_{(t-1)ljig} + \boldsymbol{\zeta}(\mathbf{X}_j y_{(t-1)ljig}) + \zeta_{1j} y_{(t-1)ljig} + \zeta_{2i} y_{(t-1)ljig} + \theta_j + \beta_i \\
&\quad + f(time_t)(group_j = g) \\
&\quad + f(trialorder_l)(group_j = g) + f(time_t, trialorder_l)(group_j = g).
\end{aligned} \tag{5}$$

The ζ_{1j} and ζ_{2i} are assumed to follow a normal distribution: $\zeta_{1j} \sim N(0, \sigma_{\zeta_1})$ and $\zeta_{2i} \sim N(0, \sigma_{\zeta_2})$.

Model C: Adding Individual-Level Trend and Learning Random Effects to Model B

The participant-specific random slopes of time and trial orders (θ_{1j} and θ_{2j}) are added to Model B to model the individual-specific deviation from the group-specific mean for time and trial orders, respectively:

$$\begin{aligned}
&\text{logit}[P(y_{tljig} = 1 | y_{(t-1)ljig}, \mathbf{X}_j, time_t, trialorder_l, \alpha_1, \boldsymbol{\alpha}, \zeta_1, \boldsymbol{\zeta}, \zeta_{1j}, \zeta_{2i}, \theta_j, \theta_{1j}, \theta_{2j}, \beta_i)] \\
&= \alpha_1 + \boldsymbol{\alpha}\mathbf{X}_j + \zeta_1 y_{(t-1)ljig} + \boldsymbol{\zeta}(\mathbf{X}_j y_{(t-1)ljig}) + \zeta_j y_{(t-1)ljig} + \zeta_i y_{(t-1)ljig} + \theta_j + \beta_i \\
&\quad + f(time_t)(group_j = g) + f(trialorder_l)(group_j = g) \\
&\quad + f(time_t, trialorder_l)(group_j = g) + \theta_{1j} time_t + \theta_{2j} trialorder_l.
\end{aligned} \tag{6}$$

The θ_{1j} and θ_{2j} are assumed to follow a normal distribution: $\theta_{1j} \sim N(0, \sigma_{\theta_1})$ and $\theta_{2j} \sim N(0, \sigma_{\theta_2})$.

3.2. Parameter Estimation

We utilize the bam function to perform the fitting of the specified models to model group-specific functional trend and learning in the mgcv package. In order to estimate the intercept difference ($\boldsymbol{\alpha}$) and group-specific smooth functions ($f(time_t)(group_j = g)$, $f(trialorder_l)(group_j = g)$, and $f(time_t, trialorder_l)(group_j = g)$) in Models A, B, and C, the group covariate ($group_j$) should be coded as factor in R. For the group-specific smooth functions, a cubic regression spline (CRS; Green & Silverman,

https://doi.org/10.1007/psy.2024.27 Published online by Cambridge University Press

1994) was selected in the mgcv. The CRS is constructed through the utilization of cubic polynomials that are interconnected at specific points known as knots. The knots are automatically distributed across the entire range of the observed covariate ($time_t$ or $trialorder_l$) at uniform intervals. When using known CRS basis functions, it is important to select an appropriate number of basis functions (K_1 and K_2). In this study, the corrected Akaike's Information Criterion (corrected AIC; Wood et al., 2016) was employed for model comparisons with differing Ks. The corrected AIC utilizes the effective degrees of freedom (*edf*) as the number of parameters necessary to represent smooth functions in the penalty term and accounts for smoothing parameter uncertainty. For the selected model regarding K_1 and K_2, the k-index (Wood, 2017) was used to check whether there are any residual patterns that the chosen K_1 and K_2 are failing to capture.

The "wiggliness" of a univariate smooth function is controlled with a quadratic smoothing penalty. As an example, the quadratic smoothing penalty for $f(time_t)(group_j = g)$ is written as:

$$\lambda_g \int_{-\infty}^{+\infty} \{f''(time_t)\}^2 d_{time} = \lambda_g \boldsymbol{\delta}_g' \mathbf{S}_g \boldsymbol{\delta}_g, \tag{7}$$

where λ_g is a smoothing parameter, $\boldsymbol{\delta}_g$ is a vector of basis coefficients, and $\mathbf{S}_g$ is a penalty matrix embedded as a diagonal block in a matrix for group g.

To measure the wiggliness of a tensor product ($f(time_t, trialorder_l)(group_j = g)$), we use a marginal penalty matrix combined with a smoothing parameter. This approach mirrors the construction of a univariate smooth function for $time_t$ and $trialorder_l$ (Wood, 2017). That is, the quadratic smoothing penalty for the tensor product is specified as follows:

$$\begin{aligned} &\int_{time,trialorder} \lambda_{time}\{f''(time_t)\}^2 + \lambda_{trialorder}\{f''(trialorder_l)\}^2 d_{time} d_{trialorder} \\ &\quad \approx \lambda_{time} \boldsymbol{\gamma}^T \tilde{\mathbf{S}}_{time} \boldsymbol{\gamma} + \lambda_{trialorder} \boldsymbol{\gamma}^T \tilde{\mathbf{S}}_{trialorder} \boldsymbol{\gamma}, \end{aligned} \tag{8}$$

where λ_{time} and $\lambda_{trialorder}$ are smoothing parameters, $\tilde{\mathbf{S}}_{time}$ is a penalty matrix defined as $\mathbf{S}_{time}{}' \otimes \mathbf{I}_{K_2'}$ ($\mathbf{S}_{time}$ is the reparameterized basis function for the approximation to the quadratic smoothing penalty, $\otimes$ is the Kronecker product, and $\mathbf{I}_{K_2'}$ is the rank $K_2{}'$ identity matrix), and $\tilde{\mathbf{S}}_{trialorder}$ is a penalty matrix calculated as $\mathbf{I}_{K_2} \otimes \mathbf{S}_{trialorder}{}'$ ($\mathbf{I}_{K_2}$ is the rank K_2 identity matrix and $\mathbf{S}_{trialorder}{}'$ is the reparameterized basis function for the approximation to the quadratic smoothing penalty).

In using mgcv package for GAMM, the random effects are treated as smooth functions with an identity matrix as penalty matrix (i.e., $\mathbf{S} = \mathbf{I}$, where $\mathbf{I}$ is an identity matrix). The variance of the random variable is the inverse of the estimated smoothing parameter: e.g., $Var(\theta_j) = (\widehat{\lambda}\mathbf{I})-1$. The predicted values of the random effects were estimated as the basis coefficients.

In the mgcv package, the smooth parameters ($\boldsymbol{\lambda}$) can be selected by either prediction error (GCV.Cp and GACV.Cp in the mgcv package) or marginal likelihood (REML and ML in the mgcv package). REML and ML are preferable to the other criteria, as they are less prone to local minima. In this study, fast restricted maximum likelihood estimation (fREML) was chosen in the bam function.

Given REML-based smoothing parameters ($\widehat{\boldsymbol{\lambda}}$) and the estimated variance matrix of the random effects (e.g., $\widehat{\Sigma} = diag(\mathbf{b} = [\theta_j, \beta_i])'$ in Eq. 1), parameters (e.g.,$\vartheta = [\boldsymbol{\alpha}, \boldsymbol{\lambda}, \boldsymbol{\zeta}]'$ in Eq. 1) are estimated using a penalized iteratively re-weighted least squares (PIRLS; Wood, 2017) with a default option, `optimizer=c("outer","newton")`, in the `mgcv` package. The following weighted least squares objective can be minimized to obtain $\widehat{\vartheta}$:

$$D(\vartheta) + \boldsymbol{\lambda}\vartheta' \mathbf{S}\vartheta + \phi \mathbf{b}' \Sigma^{-1} \mathbf{b}, \tag{9}$$

where $D(\vartheta)$ is the model deviance ($D(\vartheta) = 2\{l_{max} - l(\vartheta)\}$, where l is the log-likelihood), $\mathbf{S}$ is a matrix in which a zero block matrix is padded for the fixed effects, and the penalty matrix of smooth functions is padded for the basis coefficients, and ϕ is the scale parameter in an exponential family distribution (for binary responses, $\phi = 1$).

https://doi.org/10.1007/psy.2024.27 Published online by Cambridge University Press

In AR modeling, the treatment of the initial response variable (y_{1ljig}) is an important issue because an AR1 covariate does not exist for the initial response variable. In this study, the initial response variable (y_{1ljig}) was not modeled due to the following two reasons as used in Cho et al. (2018, 2020, 2022). First, there are a large number of time points (252 time points). For models with random effects and autoregressive responses, the issue of missing initial responses tends to be less pronounced when the number of time points is substantial (Hsiao, 2003). Second, omitting the initial two time points (t = 1 [180 ms] and t = 2 [190 ms]) is not anticipated to influence the subsequent responses. The variation between 180 ms and 220 ms (the initial 40 ms of data) is expected to be minimal. Eye-tracking studies frequently analyze fixed condition effects in a designated time window starting 200 ms after the introduction of a pivotal word. In parallel, a pre-200 ms baseline duration is typically scrutinized to identify existing data patterns before the primary analysis interval (e.g., Brown-Schmidt & Fraundorf, 2015).

3.3. Model Comparisons and Evaluation

Candidate models (Models A, B, and C) were compared based on the corrected AIC. For a selected model based on the corrected AIC, model adequacy was evaluated using residual analysis. The Pearson residual was calculated as $r_{tljig} = \{y_{tljig} - E(y_{tljig})\}/\sqrt{Var(y_{tljig})}$, where $E(y_{tljig})$ and $Var(y_{tljig})$ are the model-based mean and variance obtained from the selected model. To interpret how well the selected model explains or predicts responses for a particular trial order, person, or item, a trial-level fit (M_l), a person-level fit (W_j), and an item-level fit (Z_i) were calculated and interpreted as the mean of r_{tljig} by trial order, person, and item, respectively. In addition, the assumptions of normality for the random effects were evaluated using Q-Q plots.

3.4. Testing Smooth Functions and Plotting Differences-Between-Smooths

The null hypothesis was tested to detect trend, learning, or their interaction effects for each group at the nominal level of 0.05. For $f(time_t)(group_j = g)$ as example, the null hypothesis $H_0 : f(time_t)(group_j = g) = 0$ indicates that the smooth function is indistinguishable from zero (i.e., no trend effect) for all $time_t$ in the range of interest. Under H_0, the test statistic T_r (Wood, 2017, pp. 305–306) follows a Chi-square distribution with degrees of freedom as the rounded effective degrees of freedom (*edf*; Wood, 2013).

For group comparisons, it is important to determine whether there are significant differences in trend or learning between diagnostic groups (e.g., NC vs. TBI in the motivating example) regarding a condition (e.g., grooming movement vs. gesture in the motivating example). The interaction between the diagnostic groups and the condition was derived based on by-variable smooth functions. As the first step to probe the interaction, the *differences* in the fitted smooth function between the levels of the condition are derived for *each* diagnostic group (NC or TBI in the motivating example) by generating the posterior distribution of the smooth function (Marra and Wood, 2012). To interpret the differences between the diagnostic groups, the mean differences between the conditions were added to the differences in the fitted smooth function. For the trend effect between Condition 1 (Group 1 as a baseline or reference group; g = 1) and Condition 2 (Group 2; g = 2) in a diagnostic group as an example,

$$\begin{aligned}&\widehat{\alpha}_2 + \left\{\tilde{f}(time_t)(group_j = 2) - \tilde{f}(time_t)(group_j = 1)\right\} \\ &\quad = \widehat{\alpha}_2 + \left[\left\{\sum_{k=2}^{K_1} \widehat{\delta}_{k.2} b_{k.2}(time_t)(group_j = 2)\right\} - \left\{\sum_{k=2}^{K_1} \widehat{\delta}_{k.1} b_{k.1}(time_t)(group_j = 1)\right\}\right],\end{aligned} \tag{10}$$

where $\widehat{\delta}_{k.2}$ and $\widehat{\delta}_{k.1}$ are basis coefficient estimates for Group 2 and Group 1, respectively. In Eq. 10, summations start with 2 because of the identification constraints. To obtain the posterior distribution, one thousand replicated basis coefficients were simulated from a multivariate normal distribution (*MVN*) for each condition:

$$\boldsymbol{\delta}_2 \sim MVN\left(\widehat{\boldsymbol{\delta}}_2, \widehat{V}_{\beta_2}\right); \boldsymbol{\delta}_1 \sim MVN\left(\widehat{\boldsymbol{\delta}}_1, \widehat{V}_{\beta_1}\right), \tag{11}$$

https://doi.org/10.1007/psy.2024.27 Published online by Cambridge University Press

where $\widehat{V}_{\beta_2}$ and $\widehat{V}_{\beta_1}$ are covariance matrix of basis coefficient estimates for Group 2 and Group 1, respectively. For the post hoc comparisons of groups, the standard deviation of *differences* between the two fitted smooth functions (e.g., standard deviation of Eq. 10 across one thousand replications) is calculated to construct 95% credible bands for testing differences in the fitted smooth functions between the levels of the condition. When the credible band does not include 0, it is concluded that there are differences in trend or learning between the levels of conditions *within* each diagnostic group. Additionally, the range of time and trial orders for the differences can be identified when the credible band does not encompass 0. In the study, the mgcViz package (Fasiolo et al., 2020) is used to detect these ranges.

As the second step, significance of the differences in the fitted smooth function (i.e., Eq. 10) *between* the diagnostic groups can be inferred. When there are overlapped credible bands over time (for trend) or over trial orders (for learning) between the diagnostic groups, we can conclude that the condition effect differs by the diagnostic groups. In the presence of non-overlapped credible bands between the diagnostic groups, the ranges of time and trial orders for the differences can be detected.

4. Illustration

In this section, the specified models in Sect. 3 are applied to the empirical data set presented in Sect. 2. The data and the R code used in the illustration can be found in the Open Science Framework, https://osf.io/q3e7u/. To answer the substantive research question, a categorical group covariate was set as follows: Group 1 ($g = 1$) for NC participants in a grooming movement condition, Group 2 ($g = 2$) for NC participants in a gesture condition, Group 3 ($g = 3$) for TBI participants in a grooming movement condition, and Group 4 ($g = 4$) for TBI participants in a gesture condition. Group 1 was set as the reference or baseline group.

4.1. Exploratory Analyses of Change Processes

To characterize trend over time graphically, the logit transform of binary outcome variables (called *empirical logit*; Cox & Snell, 1989) was calculated for each participant j in group g at a time point t. To render the empirical logit unbiased, it was defined by adding $0.5/L$ to each observed proportion (e.g., see p. 32, Cox & Snell, 1989):

$$\text{emlogit}_{tjg} = \log\left(\frac{Prop_{tjg} + 0.5/L}{1 - Prop_{tjg} + 0.5/L}\right), \tag{12}$$

where $Prop_{tjg}$ is the proportion of a fixation "1" across the number of trial orders L. Adding $0.5/L$ has the further advantage of ensuring that the empirical logit is defined in cases where $Prop_{tjg} = 0$ or $Prop_{tjg} = 1$. In a similar way, the empirical logit of the proportion of a fixation '1" across the time points T for each participant j in group g at a trial order l was obtained to explore learning over trial orders graphically as follows:

$$\text{emlogit}_{ljg} = \log\left(\frac{Prop_{ljg} + 0.5/T}{1 - Prop_{ljg} + 0.5/T}\right), \tag{13}$$

where $Prop_{ljg}$ is the proportion of a fixation "1" across the number of time points T.

Figure 2 (top and middle) presents the time series plots for group-level trend and learning using emlogit_{tjg} and emlogit_{ljg}, respectively, by four groups.[3] To plot the group-level trend, the smooth function of the individual-level emlogit_{tjg} was fitted using the CRS in the `gam` function of the `mgcv`

[3]On the y axis of the time series plots (top and middle) in Fig. 2, the full ranges of emlogit_{tjg} and emlogit_{ltg} are presented to accommodate the models displaying the group-level trend and learning in the ggplot function; individual-level trend and learning were not included in the time series plots (top and middle) to maintain clarity and avoid overcrowding.

https://doi.org/10.1007/psy.2024.27 Published online by Cambridge University Press

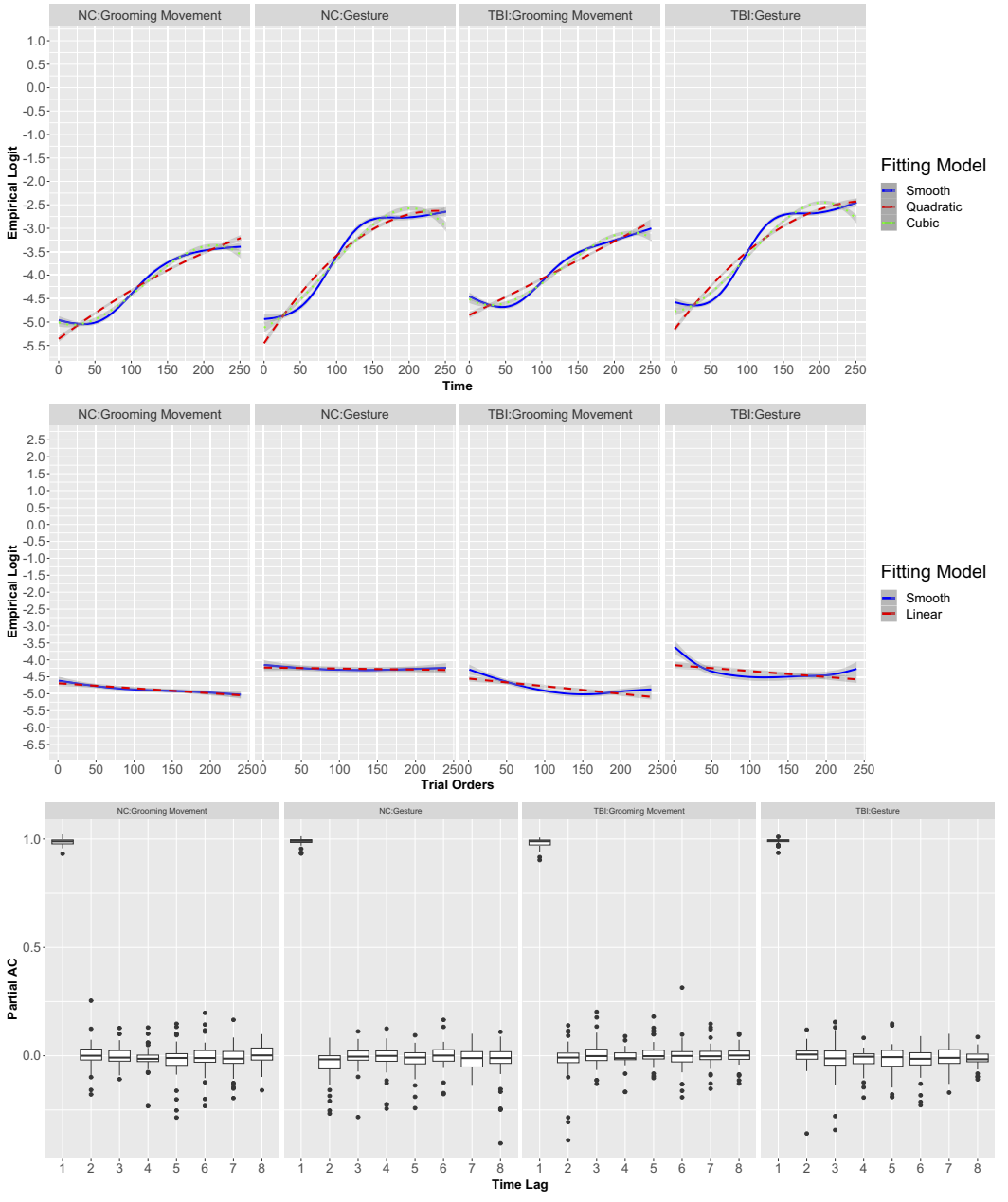

Figure 2. Empirical study: A time series plot illustrating group-level trend (top), a time series plot demonstrating group-level learning (middle), and a partial autocorrelation plot for AR1 (bottom).

package. To facilitate comparison, polynomial regressions with quadratic and cubic terms were also fitted to the same data set. It was observed that the fitted smooth functions deviated from the fitted polynomial regressions. Additionally, target fixation (on the *y* axis) was higher in the gesture condition compared to the grooming movement condition, for both the NC and TBI groups. However, there were more target fixations at earlier time points in the NC group than in the TBI group. To examine group-level learning, the smooth function of the individual-level emlogit$_{ljg}$ was fitted using the CRS

https://doi.org/10.1007/psy.2024.27 Published online by Cambridge University Press

method, and a linear regression model was also fitted to the same data set for comparison. In the NC group, a linear decreasing pattern in target fixation over trial orders was observed. However, in the TBI group, the fitted smooth function deviated from the linear line and presented a U-pattern, indicating a decreasing pattern in target fixations at the beginning followed by an increasing pattern in later trial orders. Figures in "Appendix A" illustrate individual-level trend and learning. It was observed that there is variability in trend and learning effects among individuals. In the following modeling, we will assess whether accounting for this variability leads to model-data fit improvement in detecting the group-level trend and learning effects.

A partial autocorrelation (AC) can be calculated to select the order of ACs (Chatfield, 2004). It was calculated using emlogit_{tjc} by groups. Figure 2 (bottom) shows the plot of the partial AR against the 1–8 time lags and a box plot at each time lag indicates variability in the partial AR across individuals. As depicted in Fig. 2 (bottom), the partial AC with an order of 1 is noticeably close to 1, while those with larger lags approach 0. This finding suggests that modeling autocorrelation effects only requires consideration of AR1. As shown in the box plot of AR1 in Fig. 2 (bottom), there is variability in partial AC across individuals. In the following analysis, we will assess whether this variability plays a role in detecting group-level trend and learning effects.

To summarize the exploratory investigation of the change processes, the results suggest the presence of a nonlinear trend, linear or nonlinear learning differing by diagnostic groups, and an AR1 effect that differs by groups. These change processes are modeled to investigate group-specific trend and learning effects.

4.2. Analyses

The specified models, Model A (Eq. 1), Model B (Eq. 5), and Model C (Eq. 6), were each fitted to the same empirical data set. When fitting these three models, models with different numbers of basis functions (K_1 and K_2) ranging from 3 to 12 were considered. For each model, seven basis functions ($K_1 = K_2 = 7$) were chosen for by-variable smooth functions. The $K_1 = K_2 = 7$ was chosen because the models with $K_1 = K_2 = 7$ consistently showed the lowest corrected AIC among the candidate models. Additionally, models with $K_1 = K_2 = 7$ had a k-index close to 1 for all by-variable smooth functions. Of the three models, Model C was selected based on 'the corrected AIC (refer to Table 1 for the corrected AIC values).

For the selected Model C, model adequacy was evaluated using Pearson residuals. Only 0.171% of observations exceeded 2 in the absolute value of Pearson residuals.[4] No trial orders, participants, or items exceeded 2 in the absolute value of Pearson residuals. These results indicate that Model C explains or predicts the data well. In the Q-Q plots of the predicted random effects in Model C (the Q-Q plots are shown in "Appendix B"), all quantile points were within the 95% confidence bands, with some deviations from normality for a few points for random slopes of AR1 in the lower extreme (having residuals falling slightly outside the 95% confidence bands). This suggests that the assumptions of normality for the random effects were generally met.

4.3. Results of the Selected Model

In the following, results of Model C in Table 1 are interpreted. The estimates in Table 1 are on the logit scale. The patterns of target fixations in NC participants in a grooming movement condition (Group 1) served as the baseline. The average log odds of a target fixation in Group 1 was −7.933 (SE=0.189, $p < 2e-16$), holding all other covariates constant. Significant differences in mean target fixations were found between the NC participants in the grooming movement condition (Group 1) and NC participants in the gesture condition (Group 2) (EST = 0.700, SE = 0.043, $p < 2e-16$), reflecting a

[4]The Pearson residual from a single observation, taken from a trial order, an individual, or an item, may deviate significantly from a normal distribution. Nonetheless, any observation with an absolute standardized residual value greater than 2 can be closely examined for discrepancies.

https://doi.org/10.1007/psy.2024.27 Published online by Cambridge University Press

Table 1. Empirical study: results of the specified models

	Model A	
	EST	SE
Fixed Effects		
Intercept[α_1]	**−7.652**	0.138
ylag1[ζ_1]	**10.678**	0.051
Group2[α_2]	**0.702**	0.043
Group3[α_3]	0.217	0.192
Group4[α_4]	**0.734**	0.191
ylag1:Group2[ζ_2]	**−0.291**	0.065
ylag1:Group3[ζ_3]	**0.310**	0.072
ylag1:Group4[ζ_4]	−0.119	0.065
Random Effects		
Person		
SD of θ_j[$\sqrt{\sigma_\theta}$]	0.871	
SD of ylag1[$\sqrt{\sigma_{\zeta_1}}$]	–	
SD of θ_{1j} [$\sqrt{\sigma_{\theta_1}}$]	–	
SD of θ_{2j} [$\sqrt{\sigma_{\theta_2}}$]	–	
Item		
SD of β_i[$\sqrt{\sigma_\beta}$]	0.110	
SD of ylag1[$\sqrt{\sigma_{\zeta_2}}$]	–	
	Ref.edf	T_r(*p*-value)
Smooth Functions		
s(timecoded):Group1[$f(time_t)(group_j = 1)$]	5.766	344.178(<2e-16)
s(timecoded):Group2[$f(time_t)(group_j = 2)$]	5.869	577.658(<2e-16)
s(timecoded):Group3[$f(time_t)(group_j = 3)$]	5.370	196.383(<2e-16)
s(timecoded):Group4[$f(time_t)(group_j = 4)$]	5.920	480.671(<2e-16)
s(trialorder):Group1[$f(trialorder_t)(group_j = 1)$]	3.189	15.385(0.002)
s(trialorder):Group2[$f(trialorder_t)(group_j = 2)$]	4.170	11.549(0.027)
s(trialorder):Group3[$f(trialorder_t)(group_j = 3)$]	3.424	27.896(9.39e-06)
s(trialorder):Group4[$f(trialorder_t)(group_j = 4)$]	5.561	60.667(<2e-16)
ti(timecoded,trialorder):Group1[$f(time_t, trialorder_l)(group_j = 1]$	6.164	16.345(0.013)
ti(timecoded,trialorder):Group2[$f(time_t, trialorder_l)(group_j = 2]$	5.425	8.683(0.145)
ti(timecoded,trialorder):Group3[$f(time_t, trialorder_l)(group_j = 3]$	8.473	8.137(0.463)
ti(timecoded,trialorder):Group4[$f(time_t, trialorder_l)(group_j = 4]$	6.620	5.590(0.551)
df	193.7	
LL	−59034.3	
Corrected AIC	118456.1	

(Continued)

https://doi.org/10.1007/psy.2024.27 Published online by Cambridge University Press

Table 1. Continued

	Model B	
	EST	SE
Fixed Effects		
Intercept[α_1]	**−7.918**	0.186
ylag1[ζ_1]	**11.333**	0.168
Group2[α_2]	**0.707**	0.043
Group3[α_3]	0.244	0.260
Group4[α_4]	**0.780**	0.259
ylag1:Group2[ζ_2]	**−0.346**	0.066
ylag1:Group3[ζ_3]	0.161	0.233
ylag1:Group4[ζ_4]	−0.238	0.231
Random Effects		
Person		
SD of θ_j[$\sqrt{\sigma_\theta}$]	1.194	
SD of ylag1[$\sqrt{\sigma_{\zeta_1}}$]	1.008	
SD of θ_{1j} [$\sqrt{\sigma_{\theta_1}}$]	–	
SD of θ_{2j} [$\sqrt{\sigma_{\theta_2}}$]	–	
Item		
SD of β_i[$\sqrt{\sigma_\beta}$]	0.099	
SD of ylag1[$\sqrt{\sigma_{\zeta_2}}$]	0.072	
	Ref.edf	T_r(*p*-value)
Smooth Functions		
s(timecoded):Group1[$f(time_t)(group_j = 1)$]	5.728	351.052(<2e-16)
s(timecoded):Group2[$f(time_t)(group_j = 2)$]	5.693	612.884(<2e-16)
s(timecoded):Group3[$f(time_t)(group_j = 3)$]	5.551	266.411(<2e-16)
s(timecoded):Group4[$f(time_t)(group_j = 4)$]	5.892	524.634(<2e-16)
s(trialorder):Group1[$f(trialorder_t)(group_j = 1)$]	2.944	9.379(0.020)
s(trialorder):Group2[$f(trialorder_t)(group_j = 2)$]	3.695	8.064(0.009)
s(trialorder):Group3[$f(trialorder_t)(group_j = 3)$]	3.387	29.384(3.99e-06)
s(trialorder):Group4[$f(trialorder_t)(group_j = 4)$]	5.294	41.995(<2e-16)
ti(timecoded,trialorder):Group1[$f(time_t, trialorder_l)(group_j = 1]$	7.091	18.751(0.010)
ti(timecoded,trialorder):Group2[$f(time_t, trialorder_l)(group_j = 2]$	11.651	19.147(0.073)
ti(timecoded,trialorder):Group3[$f(time_t, trialorder_l)(group_j = 3]$	9.536	12.213(0.235)
ti(timecoded,trialorder):Group4[$f(time_t, trialorder_l)(group_j = 4]$	10.048	12.222(0.278)
df	284.2	
LL	−58256.1	
Corrected AIC	117080.6	

(Continued)

https://doi.org/10.1007/psy.2024.27 Published online by Cambridge University Press

Table 1. Continued

	Model C	
	EST	SE
Fixed Effects		
Intercept[α_1]	**−7.933**	0.189
ylag1[ζ_1]	**11.289**	0.168
Group2[α_2]	**0.700**	0.043
Group3[α_3]	0.233	0.264
Group4[α_4]	**0.776**	0.263
ylag1:Group2[ζ_2]	**−0.363**	0.066
ylag1:Group3[ζ_3]	0.153	0.234
ylag1:Group4[ζ_4]	−0.251	0.232
Random Effects		
Person		
SD of θ_j[$\sqrt{\sigma_\theta}$]	1.136	
SD of ylag1[$\sqrt{\sigma_{\zeta_1}}$]	1.011	
SD of θ_{1j} [$\sqrt{\sigma_{\theta_1}}$]	0.002	
SD of θ_{2j} [$\sqrt{\sigma_{\theta_2}}$]	0.003	
Item		
SD of β_i[$\sqrt{\sigma_\beta}$]	0.098	
SD of ylag1[$\sqrt{\sigma_{\zeta_2}}$]	0.046	
	Ref.edf	T_r(*p*-value)
Smooth Functions		
s(timecoded):Group1[$f(time_t)(group_j = 1)$]	5.781	245.300(<2e-16)
s(timecoded):Group2[$f(time_t)(group_j = 2)$]	5.896	446.700(<2e-16)
s(timecoded):Group3[$f(time_t)(group_j = 3)$]	5.489	184.900(<2e-16)
s(timecoded):Group4[$f(time_t)(group_j = 4)$]	5.943	397.400(<2e-16)
s(trialorder):Group1[$f(trialorder_t)(group_j = 1)$]	1.931	11.040(0.006)
s(trialorder):Group2[$f(trialorder_t)(group_j = 2)$]	3.250	7.397(0.082)
s(trialorder):Group3[$f(trialorder_t)(group_j = 3)$]	3.262	25.640(1.96e-05)
s(trialorder):Group4[$f(trialorder_t)(group_j = 4)$]	5.176	40.330(1.69e-06)
ti(timecoded,trialorder):Group1[$f(time_t, trialorder_l)(group_j = 1]$	5.451	17.490(0.006)
ti(timecoded,trialorder):Group2[$f(time_t, trialorder_l)(group_j = 2]$	6.316	12.710(0.057)
ti(timecoded,trialorder):Group3[$f(time_t, trialorder_l)(group_j = 3]$	8.947	10.590(0.298)
ti(timecoded,trialorder):Group4[$f(time_t, trialorder_l)(group_j = 4]$	5.699	7.565(0.241)
df	378.3	
LL	−58026.4228	
Corrected AIC	116809.5	

-indicates that an effect or a smooth function is not modelled; Significance for fixed effects is presented in bold at .05; LL indicates a log-likelihood value; NumP indicates the number of parameters; Group 1 ($g = 1$) is for NC participants in a grooming movement condition, Group 2 ($g = 2$) is for NC participants in a gesture condition, Group 3 ($g = 3$) is for TBI participants in a grooming movement condition, and Group 4 ($g = 4$) is for TBI participants in a gesture condition

https://doi.org/10.1007/psy.2024.27 Published online by Cambridge University Press

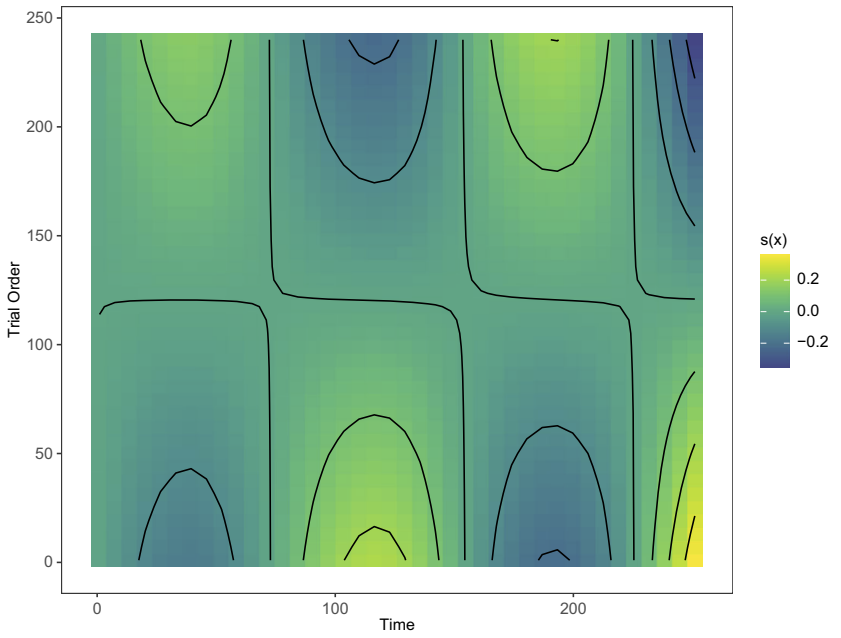

Figure 3. Empirical study: Predicted two-dimensional by-variable smooth function of $(time_t, trialorder_l)$. *Note.* In the figure, lines in the contour plot are drawn connecting the $(time_t, trialorder_l)$ coordinates where the same value (the logit transform of probability) occurs, with the effect of the interaction increasing as the color of the contour plot becomes warmer (from blue to yellow); no credible bands were added to avoid cluttered figures (Color figure online).

benefit of gesture in the NC group. In addition, there was a significant difference in mean target fixations between NC participants in the grooming movement condition (Group 1) and TBI participants in the gesture condition (Group 4, EST = 0.776, SE = 0.263, $p = 0.003$), showing a benefit of gesture in the TBI group compared to the NC group in the grooming condition. Furthermore, there was no significant mean difference between the NC participants the a grooming movement condition (Group 1) and TBI participants in the grooming movement condition (Group 3) (EST=0.233, SE=0.264, $p = 0.379$). These findings indicate that in the absence of a helpful gesture, that participants with and without TBI identified and fixated the target referent at a similar rate on average, and that in addition the gesture was helpful in directing participants with and without TBI toward the target.

As shown in Table 1, the two-dimensional by-variable smooth function of $(time_t, trialorder_l)(group_j = g)$ was significant only for the NC participants in the grooming movement condition (Group 1). This significant interaction between time and trial orders indicates that the trend over time within a trial changed across trial orders of the experiment. Figure 3 displays the predicted two-dimensional by-variable smooth function for Group 1. In the figure, higher target fixations are evident toward the middle and end of the time variable within the [1, 125] trial order range. In contrast, increased fixations appear at the beginning and later-middle sections of the time variable in the [126, 240] range. The pattern in the early trials of the experiment may be due to participants fixating the video of the actor (at the expense of target fixations) at the beginning of the trial (corresponding to the moment the speaker is producing the grooming movement), then alternating between the target and video thereafter. At these early trials in the experiment participants may still be learning that grooming movements are not providing an informative cue, and therefore still be looking to the video initially. As trial order increases in the [126, 240] range, participants may be less likely to gaze at the video initially as they may have learned that the video is not informative, which allows for serendipitous early target fixations.

In addition, the by-variable smooth functions of $time_t$ and $trialorder_l$ were significant for all four groups, except for the by-variable smooth function of $trialorder_l$ of the NC participants in the gesture

https://doi.org/10.1007/psy.2024.27 Published online by Cambridge University Press

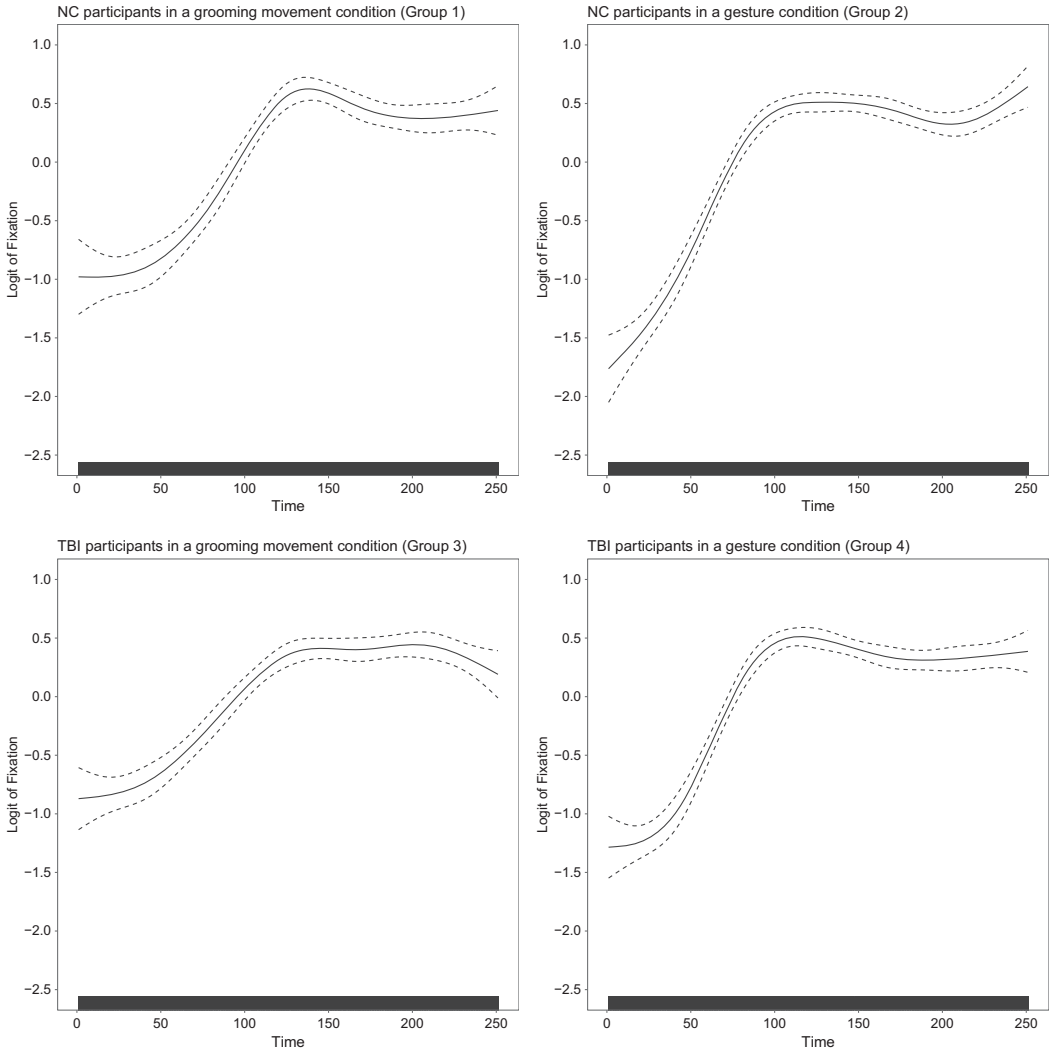

Figure 4. Empirical study: Predicted by-variable smooth function of $time_t$ for Group 1 (top, left), Group 2 (top, right), Group 3 (bottom, left), and Group 4 (bottom, right). *Note.* The dotted lines indicate the 95% credible bands of the predicted values; each tick mark in the *x*-axis represents time points.

condition (Group 2). Figures 4 and 5 present the predicted by-variable smooth functions of $time_t$ and $trialorder_l$, respectively. As shown in Fig. 4, all four groups show a generally increasing pattern of target fixations across time within the trial that generally plateaus at later time points. The initial increase observed between time points 0 and 150 reflects the interpretation of the verb and gesture (in Groups 2 and 4). The plateau after time point 150 likely reflects the fact that many participants have identified the target by this time point. Across trial orders, three of the four groups (all but Group 2) showed generally decreasing target fixations across trial orders within the experiment. We speculate that this decreasing target fixations may reflect either fatigue, or possibly increased efficiency in identifying the target across trial orders. If participants learn to quickly identify the target, it may require fewer or shorter fixations; if so, this could explain the general decrease across trial orders in target fixations. Lastly, random variability in deviations from functional group differences across individuals was small,

https://doi.org/10.1007/psy.2024.27 Published online by Cambridge University Press

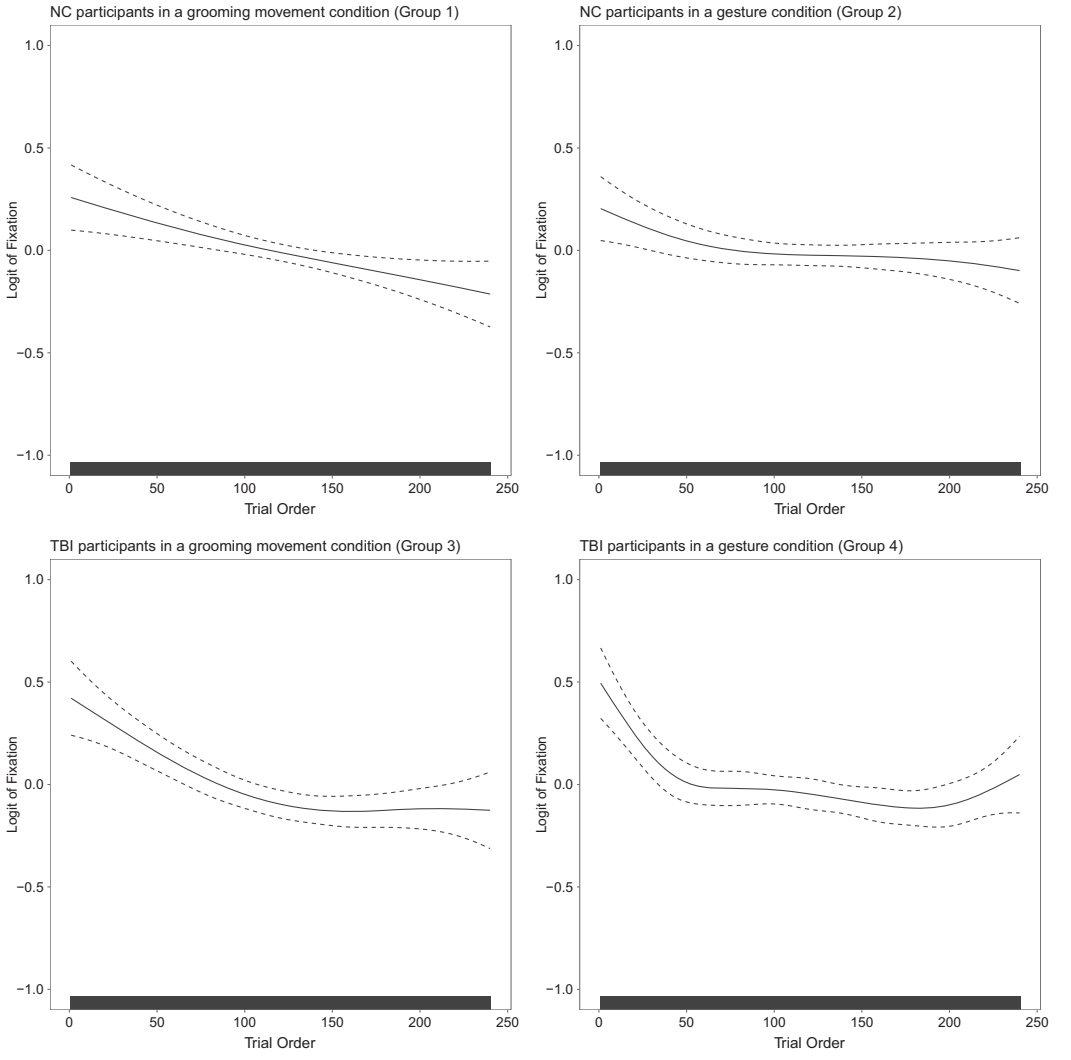

Figure 5. Empirical study: Predicted by-variable smooth function of $trialorder_l$ for Group 1 (top, left), Group 2 (top, right), Group 3 (bottom, left), and Group 4 (bottom, right). *Note.* The dotted lines indicate the 95% credible bands of the predicted values; each tick mark in the *x*-axis represents trial orders.

as evidenced by the small standard deviations of random slopes for $time_t$ and $trialorder_l$ (0.002 and 0.003, respectively).

To probe the interaction between diagnostic groups (NC vs. TBI) and conditions (grooming movement vs. gesture) over time, differences between the two condition levels were derived from the predicted by-variable smooth functions of $time_t$ and the fixed group difference effects for each diagnostic group (as shown in Eq. 10). Figure 6 (top) displays the differences (represented by lines) along with their 95% credible bands (dotted lines) for each diagnostic group over time. We infer significance from intervals that do not encompass 0.[5] As depicted in Fig. 6 (top), significant differences between the two condition levels (gesture–grooming movement) emerged in the time course of [21, 251] for the

[5]In this setting, the two diagnostic groups (NC vs. TBI) were compared. Thus, the Type I error was not controlled for the comparison.

https://doi.org/10.1007/psy.2024.27 Published online by Cambridge University Press

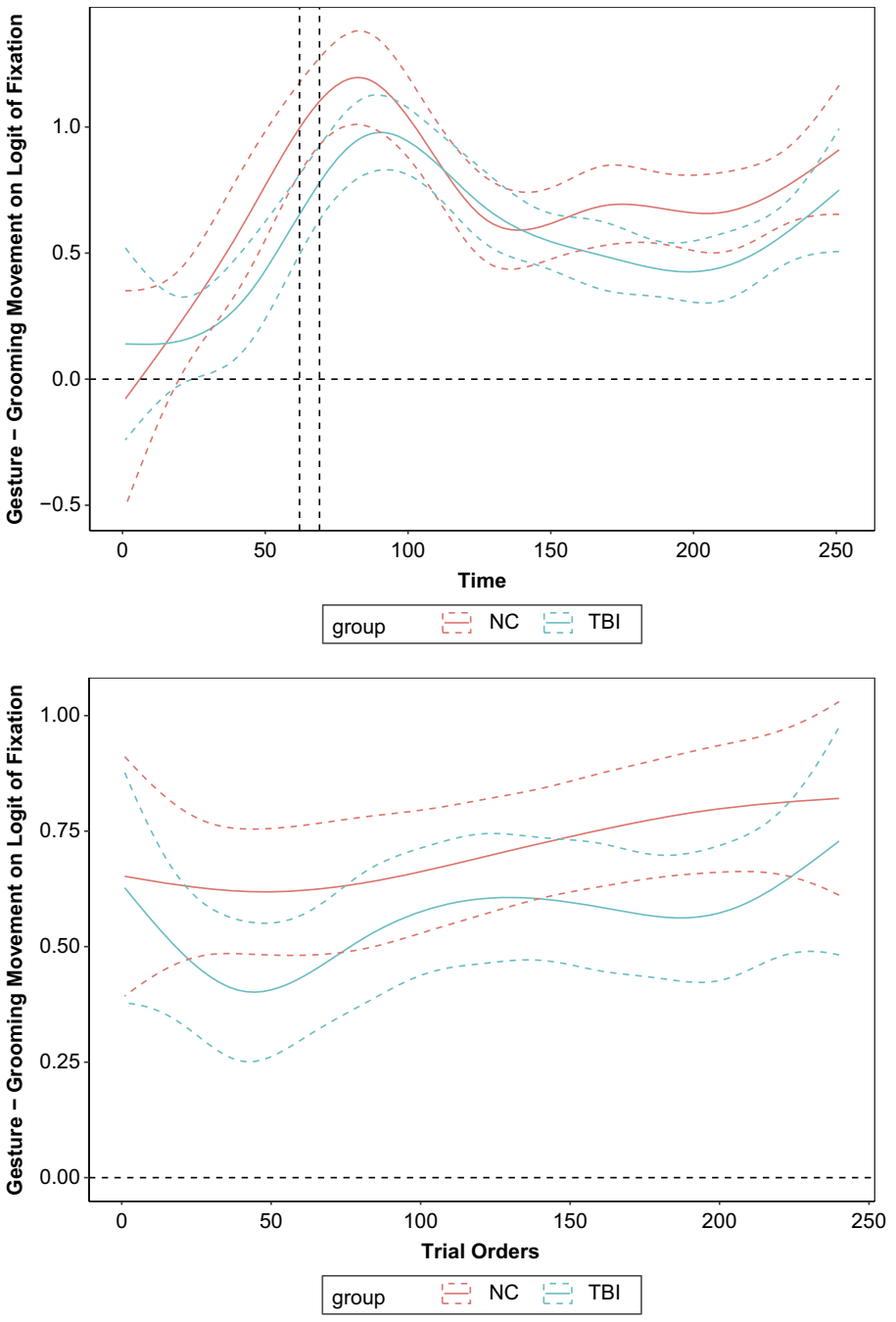

Figure 6. Empirical study: Differences in trend (top) and learning (bottom) between grooming movement and gesture conditions by NC versus TBI. *Note.* The dotted lines indicate the 95% credible bands of the fitted values; dotted vertical lines in the differences in trend (top) indicate the range of significance for the diagnostic group differences.

NC group and in the time course of [26, 251] for the TBI group (indicated by the fact that the 95% credible bands did not encompass 0 in these ranges). In both diagnostic groups, a similar pattern in the differences between the two experimental conditions (gesture–grooming movement) was observed:

https://doi.org/10.1007/psy.2024.27 Published online by Cambridge University Press

the differences were large, then they decreased, only to increase again toward the end. However, larger differences were found in the earlier time course in the NC group compared to the TBI group, which suggests that participants in the NC group were better able to take advantage of the meaningful gesture cue early in the trial compared to the TBI group. In addition, the differences between the two conditions differed *between* the diagnostic groups for the time range of [62, 69] (presented with dotted vertical lines in Fig. 6 [top]), as indicated by non-overlapped 95% credible bands. This result suggests that participants in the NC group benefitted from the gesture cue in terms of identifying the target more than participants with TBI in the short time range of [62, 69].

Similarly, to further explore the interaction between the diagnostic groups (NC vs. TBI) and conditions (grooming movement vs. gesture) over trial orders, *differences* between the two condition levels were derived. These differences were obtained from the predicted by-variable smooth functions of $trialorder_l$ combined with the fixed group difference effects for each diagnostic group. As depicted in Fig. 6 (bottom), significant differences between the two condition levels (gesture–grooming movement) emerged over trial orders in both diagnostic groups. This is indicated by the fact that the 95% credible bands did not encompass 0. The differences between the two condition levels generally increased over successive trial orders in the NC group. This suggests that the benefit of the gesture over the grooming movement grew over time across trials for NC participants—this finding is consistent with the fact that NC participants *learned* to take advantage of the gesture across trials. In contrast, these differences fluctuated within the TBI group, where an increasing trend was noted between the 50th and 140th trial orders. This fluctuation may reflect less stability in use of gesture information in the TBI group across trials or increased fatigue diminishing the benefit of gesture at later trials in the experiment. Furthermore, the change over trials in the gesture versus grooming benefit did not differ between NC and TB groups, as evidenced by overlapping 95% credible bands across trial orders.

Estimates of AR1 and standard deviation of random slopes for AR1 in Model 3 in Table 1 are interpreted below. The mean AR1 effect for NC participants in the grooming movement condition (Group 1) was 11.289 (SE=0.168, $p < 2e-16$), which is the log-odds ratio for the current response due to the previous response changing from 0 (non-fixation) to 1 (fixation) (holding all other covariates constant). These large effects of the AR1 (*exp* (11.289) log-odds ratio) indicate that there are strong carryover effects: from target at time point $t-1$ to target at time point t, and from non-target at time point $t-1$ to non-target at time point t. Significant differences in the mean AR1 effect were found only between the NC participants in the grooming movement condition (Group 1) and NC participants in the gesture condition (Group 2) (EST=−0.363, SE=0.066, $p = 3.72e-08$). The attenuated AR1 in Group 2 may reflect greater potential for change over time in this group as there was more information in the visuo-linguistic signal to drive changes in fixations. The standard deviations of the random slope of AR1 across persons and items were 1.011 and 0.046, respectively, which indicates that there is non-ignorable variability in AR1 mainly across participants.

5. Simulation Study

A simulation study was designed to demonstrate accuracy of parameter estimates and predictions of the selected model implemented in the magv package.

5.1. Simulation Design and Analysis

For the parameter recovery study, the estimates from Model 3 in Table 1 were taken as the true parameters. The covariates from the illustrations were then used to generate five hundred data sets. Bias and root mean square error (RMSE) were obtained to evaluate the accuracy of the parameter estimates. Furthermore, the mean standard error estimates (M(SE)) of fixed effects across five hundred replications were compared with the standard deviations (SD) of the estimates of fixed effects to evaluate the accuracy of standard error (SE) estimates.

https://doi.org/10.1007/psy.2024.27 Published online by Cambridge University Press

To evaluate the accuracy of the regenerated smooth functions, the root mean prediction error (RMPE) was calculated by comparing predicted values (derived from basis coefficient estimates) to true values (based on the basis coefficient parameters). For instance, in the case of functional trend effect of Group 1 ($g = 1$), RMPE can be calculated as $\sqrt{\{\sum_{k=2} K_1 \widehat{\delta}_{k.1} b_{k.1}(time_t)(group_j = 1) - f(time_t)(group_j = 1)\}2/N}$, where N is the number of observations. The RMD is interpreted as the standard deviation of the differences between predicted and the generated by-variable smooth functions.

There were no convergence problems in any simulation replications. With $K_1 = K_2 = 7$ (used in data generation), the k-index was close to 1 for smooth functions and the corrected AIC was the smallest for a model with $K_1 = K_2 = 7$ among candidate models with $K_1 = K_2 = 5,7,9$ for all replications.

5.2. Simulation Results

Table 2 presents the results of parameter recovery and RMPE of by-variable smooth functions. The biases in the fixed effect estimates and in the SDs of random effects were nearly zero. The RMSEs of these estimates were comparable to those of the estimates of GAMM for binary data (e.g., Cho et al., 2022). In addition, the ratio of M(SE) to SD was close to 1 for the fixed effect estimates, which indicates that SEs are approximately correct. For each smooth function by variable, the average RMPE across 500 replications is presented in Table 2. The average RMPE of 12 by-variable smooth functions ranged from 0.004 to 0.019, which suggests that the predicted smooth functions are close to the generated smooth functions. Taking all results together, parameters and by-variable smooth functions of Model 3 were recovered well.

6. Summary and Discussion

In this study, a model specification was provided to model group-specific functional trend and learning effects using by-variable smooth functions of time and trial orders. The model specification was formulated as GAMM, which allowed for the use of the freely available mgcv package in R. The model specification was motivated and applied to explore the differences in real-time language comprehension abilities between individuals with and without brain injury, and how these abilities facilitate learning over time. The empirical study showed that the by-variable smooth functions allowed for a numerical evaluation of a curve across the entire span of time and trial orders. This in turn facilitated the examination of subtle, yet significant, patterns in trend and learning. In addition, methods to test differences-between-smooths plots were provided and illustrated to detect differences in trend and trial orders within and between diagnostic groups and conditions. Furthermore, the parameters of specified models were accurately recovered, and the by-variable smooth functions were adequately predicted under simulation conditions that mirrored the empirical study, in using the mgcv package.

6.1. What Did We Learn from the Specified Models?

We went into this project hypothesizing that individuals with TBI may experience disruptions in processing language in real time and in context, and in integrating the unfolding linguistic signal with gesture to derive meaning. We also speculated that TBI may impair the ability to learn from language processing experiences in the moment, leading to deficits in learning over time. The results of Model 3 revealed that non-injured comparison participants benefitted from the informative gesture cue, making more target (e.g., "sandwich") fixations as they interpreted the unfolding sentence, e.g., "She will eat the very good…" when the speaker in the video produced a sandwich-eating gesture around the time of the verb, in comparison to a grooming gesture. Participants with TBI did not differ from non-injured comparison participants in the overall level of target fixations when processing sentences in the absence

https://doi.org/10.1007/psy.2024.27 Published online by Cambridge University Press

Table 2. Simulation study: results of fixed and random effects (top) and RMPE of by-variable smooth functions (bottom)

	Bias	RMSE	Ratio(M(SE)/SD)
Fixed Effects			
Intercept[α_1]	0.032	0.136	1.075(0.142/0.132)
ylag1[ζ_1]	−0.032	0.060	0.973(0.050/0.051)
Group2[α_2]	−0.035	0.055	1.017(0.043/0.042)
Group3[α_3]	−0.033	0.202	0.996(0.198/0.199)
Group4[α_4]	−0.029	0.200	0.996(0.197/0.198)
ylag1:Group2[ζ_2]	0.003	0.065	0.982(0.064/0.065)
ylag1:Group3[ζ_3]	−0.007	0.071	0.976(0.069/0.071)
ylag1:Group4[ζ_4]	0.001	0.064	0.987(0.063/0.064)
Random Effects			
Person			
SD of θ_j[$\sqrt{\sigma_\theta}$]	−0.055	0.167	
SD of ylag1[$\sqrt{\sigma_{\zeta_1}}$]	0.040	0.081	
SD of θ_{1j} [$\sqrt{\sigma_{\theta_1}}$]	−0.001	0.001	
SD of θ_{2j} [$\sqrt{\sigma_{\theta_2}}$]	−0.001	0.001	
Item			
SD of β_i[$\sqrt{\sigma_\beta}$]	−0.030	0.045	
SD of ylag1[$\sqrt{\sigma_{\zeta_2}}$]	0.051	0.091	
Smooth Functions	RMPE		
s(timecoded):Group1[$f(time_t)(group_j = 1)$]	0.011		
s(timecoded):Group2[$f(time_t)(group_j = 2)$]	0.010		
s(timecoded):Group3[$f(time_t)(group_j = 3)$]	0.011		
s(timecoded):Group4[$f(time_t)(group_j = 4)$]	0.008		
s(trialorder):Group1[$f(trialorder_t)(group_j = 1)$]	0.007		
s(trialorder):Group2[$f(trialorder_t)(group_j = 2)$]	0.007		
s(trialorder):Group3[$f(trialorder_t)(group_j = 3)$]	0.019		
s(trialorder):Group4[$f(trialorder_t)(group_j = 4)$]	0.011		
ti(timecoded,trialorder):Group1[$f(time_t,trialorder_l)(group_j = 1]$	0.014		
ti(timecoded,trialorder):Group2[$f(time_t,trialorder_l)(group_j = 2]$	0.004		
ti(timecoded,trialorder):Group3[$f(time_t,trialorder_l)(group_j = 3]$	0.008		
ti(timecoded,trialorder):Group4[$f(time_t,trialorder_l)(group_j = 4]$	0.007		

of helpful gesture. Participants with TBI *did* show a benefit from gesture compared to non-injured participants without the gesture.

Other aspects of the findings point to differences in how language processing unfolded in the two groups. Non-injured participants had significantly smaller AR1 in the gesture condition compared to the grooming condition, which may indicate greater potential for responsivity in gaze with gesture in the NC group. The AR1 effect did not differ between the NC participants in the grooming condition

https://doi.org/10.1007/psy.2024.27 Published online by Cambridge University Press

(Group 1) and participants with TBI in either the grooming or gesture conditions (Groups 3 and 4). Both participant groups in both conditions showed an increasing pattern of target fixations across time within the trial that generally plateaued at later time points in the trial. This temporal effect points to similarity across groups in processing the speech (and gesture when available) over time within the trial. Across trial orders, three of the four groups (all but Group 2—NC participants in the gesture condition) showed generally decreasing target fixations across trials within the experiment. This finding was a surprise, as we expected that there would be more target fixations at later trials, due to greater learning. Instead, this could reflect fatigue, or possibly increased efficiency in later trials. In this study, the participants experienced each item multiple times across trials of the experiment, so upon hearing the verb, e.g., "eat" they may have quickly identified the two possible targets ("sandwich" and "apple"). Exploring whether fatigue or efficiency (potentially in tandem with learning) explains changes across trials remains a question for future work.

Lastly, direct comparisons between grooming and gesture conditions across time within and across trials in the two groups revealed that the benefit of the informative gesture over the uninformative grooming movement grew over time both within and across trials for NC participants. Within a trial, NC participants benefitted more than TBI participants from the gesture, particularly early in the trial. Across trials in the experiment, NC participants showed greater gains in target fixations, compared to TBI participants. In the context of an overall decreasing trend for target fixations across trials (except for NC participants with gesture), these findings suggest that NC participants may have benefitted more from gesture in the moment (particularly early in the trial as they processed the verb) and then NC participants showed a greater propensity to learn to take advantage of the gesture across trials. In contrast, this advantage for gesture over grooming movements was less stable and fluctuated within the TBI group. These findings point to the importance of considering not just overall changes across time when learning can be expected but how different groups may *differently* learn to take advantage of useful cues in the environment.

6.2. Discussion and Methodological Limitations of the Current Study

In this study, we considered the functional group trend and learning effects using by-variable smooth functions and random effects to account for individual deviations from the group effects (i.e., the population mean). In our empirical analysis, we demonstrated that the random effects (i.e., the individual-specific deviation from the group-specific mean) can be considered based on checks for normality and model-data fit. For other empirical data sets, both the population mean and individual-specific trend and learning effects can be modeled using smooth functions (e.g., Durbán et al. (2005) for intensive continuous longitudinal data). It is feasible to fit the model, accounting for both functional group and individual trends and learning effects, using the mgcv package. Example code in R can be found in the Open Science Framework, https://osf.io/q3e7u/. However, modeling these effects, especially in intensive binary longitudinal data, may lead to computational challenges.

In PIRLS, there is no need for marginalization of random effects, leading to computational efficiency when dealing with intensive binary longitudinal data sets. However, a limitation of the PIRLS estimation method is that it does not allow for estimating correlations between random intercepts and random slopes. When an alternative estimation method becomes available, further studies are needed to assess the consequences of the limitation. This can be done by comparing parameter recovery across different estimation methods.

In line with intensive binary eye-tracking studies (e.g., Cho et al., 2018, 2020), AR1 effects were pronounced in the empirical data set. This is attributed to the serial correlation within a trial, inherent to the way our eyes move (fixations and saccades). In addition, there were non-ignorable trend effects over time within a trial in the empirical data set. There are two methodological issues in these non-stationary time series data with strong AR effects.[6] The first issue concerns the interpretation of trend

[6]The first author thanks Dr. Gregory Camilli for addressing these issues.

https://doi.org/10.1007/psy.2024.27 Published online by Cambridge University Press

and AR effects in non-stationary time series data. It is often recommended to de-trend time series data by employing a difference parameter to allow for a more accurate interpretation of AR effects. This is because trend is considered to be random variation (noise) rather than systematic change in some applications (e.g., Box et al., 2008). However, de-trending introduces AR effects and does not allow for modeling trend effects that are of interest (see Huitema 2011, pp. 414–424 for an example). Without de-trending, both trend and AR effects have been modeled simultaneously in non-stationary binary time series data (e.g., Gao et al., 2018; Kedem, & Fokianos, 2002). A challenge with having trend and AR effects simultaneously is that the trend can be interpreted through AR effects. In the current study, our interpretation is that the AR1 effect primarily captures the short-term dependencies in the data (around the trend), while the trend represents the long-term direction or pattern.

The second issue is that the use of lagged response variables (e.g., $y_{(t-1)ljig}$ in this study) can artificially *suppress* the effects of other covariates in continuous time series data when using ordinary least squares (Achen, 2000). This means that the effects of covariates of interest (e.g., trend and experimental condition effects) can be underestimated in the presence of the AR effects of the lagged response variable in a model. However, Keele and Kelly (2006) showed that the use of the lagged response variable remains appropriate in stationary continuous time series data when the past matters for the current values of the process being studied (a dynamic study), as is the case in eye-tracking data. Yet, for binary time series data, the challenges associated with lagged response variables remain less well explored. To be best of our knowledge, the most widely used currently available approach is to include the lagged response variable to account for AR in the binary time series data (e.g., Cox & Snell, 1989; Diggle et al., 1994; Fokianos & Kedem, 2003; Zeger & Qaqish, 1988); and alternative methods for modeling AR effects in binary time series data are still in the nascent stages. In the present study, we considered AR effects because excluding them leads to biased estimates and standard errors of fixed effects of interest (Cho et al., 2018). In addition, including the lagged response variables is theoretically appropriate due to the fact that the eyes tend to move in a rapid, ballistic fashion, and then linger on behaviorally relevant interest areas during the period of fixation, features which are captured in detail due to the high-resolution of modern eye-trackers. Kedem and Fokianos (2002) showed via a simulation study that the estimates and standard errors of trend and AR1 from the lagged response variable in a logistic regression model were accurately estimated in the non-stationary binary time series data having time points of 200, 300, and 1000. In addition, recent findings by Wang et al. (2023) from a simulation study suggested that the trend effect in GLMM remained unbiased in the presence of strong AR1 effects from the lagged response variable in non-stationary binary time series data, as long as the data have a substantial number of time points (greater than 200).

In psycholinguistics, the empirical logit transformation of the proportion over a series of binned time points, based on the number of observations per bin, has been utilized for continuous time series modeling from the visual world paradigm (e.g., Porretta et al., 2017; Ito & Knoeferle, 2022). In such modeling, AR effects can be incorporated into the residuals of the continuous time series data (e.g., Baayen et al., 2017). Consequently, this allows for the separation of trends from AR effects, enabling direct interpretation of the trend. However, unlike the case with reaction time (Baayen et al., 2017, 2018; Chuang et al., 2021), modeling the empirical logit of eye-tracking data in a continuous time framework presents challenges. As Porretta et al. (2017) noted, the number of observations per bin for the empirical logit transformation is inherently linked to the sampling rate of eye-trackers and the size of the bin. Moreover, an adjustment factor (typically, 0.5, Ito & Knoeferle, 2022) must be added to the proportion to render it unbiased and to prevent the return of infinite values. In addition, applying the empirical logit transformation to proportions often leads to a significantly non-normal distribution of data, unless the temporal window of aggregation is quite large. This stems from the pronounced AR process in the eye-tracking data—with small temporal windows, the empirical logit is applied to ratios such as 1/0 or 0/1, in addition to an adjustment factor included to render it unbiased and to prevent infinite values. While employing larger temporal windows may increase normality, doing so would hinder our ability to detect the dynamic signals that make eye-tracking data valuable.

https://doi.org/10.1007/psy.2024.27 Published online by Cambridge University Press

To summarize, the two approaches to model AR processes have their respective strengths and weaknesses: (a) the use of a lagged response variable for binary responses and (b) the incorporation of AR effects in the residuals using the empirical logit. Comparing the relative performance of these methods in capturing change processes (both trend and AR) in eye-tracking data is beyond the scope of the current study. In sum, the present study contributes to one of several methods for modeling AR in binary time series data.

While our simulation study results indicated that the mgcv package can accurately estimate parameters and standard errors, and predict by-variable smooth functions under conditions that mimic the empirical study as a case study, further research is needed to extend these findings to other data structures. These include varying (a) the number of time points, trial orders, persons, and items; (b) the magnitudes of effects; and (c) the shapes of by-variable smooth functions.

6.3. Broad Impact of the Current Study

A key unanswered question regarding the origins of cognitive-communicative deficits following brain injury is how these deficits arise in the moment as communication is ongoing, and if and how those deficits accrue to generate deficits over larger time scales. In the present work we examine the impact of moderate–severe TBI on the online processing of gesture-accompanied speech in context, and changes in processing over larger time scales. The modeling techniques used here revealed that participants with TBI do significantly benefit from informative gesture when processing speech. This finding points to the potential utility of gesture when communicating with individuals with cognitive-communicative impairment. However, detailed analyses of the first several hundred milliseconds of language processing, as well as analyses of how processing changed over tens of minutes within the experiment revealed new insights into the nature of the processing deficits in TBI. Of particular interest was the novel finding that during online processing within a trial, participants with TBI did not benefit from gesture as much as non-injured participants, and as the experiment progressed, participants with TBI showed weaker gains in the use of gesture. These findings tentatively suggest that deficits in capitalizing on meaningful communicative cues within rich communicative contexts in TBI may weaken the ability to learn to take advantage of those cues across time. These findings point to the importance of considering how different groups may process informative cues differently and, in turn, differ in the ability to learn to take advantage of those cues more efficiently over time.

Open science statement. The data and the R code used in the illustration can be found in the Open Science Framework, https://osf.io/q3e7u/.

Funding statement. This work was supported by NIDCD F31 DC020388 and CAPCSD Ph.D. Scholarship awarded to Clough and NIDCD Grant R01 NIH DC017926 awarded to Duff and Brown-Schmidt. Nothing in this article necessarily reflects the positions or policies of the agency, and no endorsement by it should be inferred.

Competing interests. The authors declare that they have no conflict of interest.

References

Achen, C. (2000). *Why lagged dependent variables can suppress the explanatory power of other independent variables*. In: Presented at the annual meeting of political methodology, Los Angeles.

Akhavan, N., Blumenfeld, H. K., Shapiro, L., & Love, T. (2023). Using lexical semantic cues to mitigate interference effects during real-time sentence processing in aphasia. *Journal of Neurolinguistics*, *68*, 101–159. https://doi.org/10.1016/j.jneuroling.2023.101159

Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*, 247–264. https://doi.org/10.1016/S0010-0277(99)00059-1

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412. https://doi.org/10.1016/j.jml.2007.12.005

https://doi.org/10.1007/psy.2024.27 Published online by Cambridge University Press

Baayen, R. H., Fasiolo, M., Wood, S., & Chuang, Y.-Y. (2022). A note on the modeling of the effects of experimental time in psycholinguistic experiments. *The Mental Lexicon*, *17*, 178–212.

Baayen, R. H., van Rij, J., De Cat, C., & Wood, S. N. (2018). Autocorrelated errors in experimental data in the language sciences: Some solutions offered by Generalized Additive MixedModels. In D. Speelman, K. Heylen, & D. Geeraerts (Eds.), *Mixed effects regression models in linguistics* (pp. 49–69). Springer.

Baayen, R. H., Vasishth, S., Kliegl, R., & Bates, D. (2017). The cave of shadows. Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language*, *94*, 206–234. https://doi.org/10.1016/j.jml.2016.11.006

Baker, C., & Love, T. (2021). It's about time! Time as a parameter for lexical and syntactic processing: An eyetracking-while-listening investigation. *Language, Cognition and Neuroscience*. https://doi.org/10.1080/23273798.2021.1941147

Box, G. E. P., Jenkins, G., & Reinsel, G. C. (2008). *Time series analysis: Forecasting and control* (4th ed.). Wiley.

Brown, M., Salverda, A. P., Dilley, L. C., & Tanenhaus, M. K. (2011). Expectations from preceding prosody influence segmentation in online sentence processing. ,*Psychonomic Bulletin&Review 18*, 1189–1196. https://doi.org/10.3758/s13423-011-0167-9

Brown-Schmidt, S., & Fraundorf, S. H. (2015). Interpretation of informational questions modulated by joint knowledge and intonational contours. *Journal of Memory and Language*, *84*, 49–74. https://doi.org/10.1016/j.jml.2015.05.002

Chatfield, C. (2004). *The analysis of time series: An introduction* (6th ed.). Chapman and Hall/CRC.

Cho, S.-J., Brown-Schmidt, S., De Boeck, P., & Naveiras, M. (2022). Space-time modeling of intensive binary time series eye-tracking data using a generalized additive logistic model. *Psychological Methods*, *27*, 307–346. https://doi.org/10.1037/met0000444

Cho, S.-J., Brown-Schmidt, S., De Boeck, P., & Shen, J. (2020). Modeling intensive polytomous time series eye tracking data: A dynamic tree-based item response model. *Psychometrika*, *85*, 154–184. https://doi.org/10.1007/s11336-020-09694-6

Cho, S.-J., Brown-Schmidt, S., & Lee, W.-Y. (2018). Autoregressive generalized linear mixed effect models with crossed random effects: An application to intensive binary time-series eye tracking data. *Psychometrika*, *83*, 751–771. https://doi.org/10.1007/s11336-018-9604-2

Chuang, Y.-Y., Fon, J., Papakyritsis, I., & Baayen, R. H. (2021). Analyzing phonetic data with generalized additive mixed models. In M. J. Ball (Ed.), *Handbook of clinical phonetics* (pp. 108–138). Routledge.

Clough, S., Brown-Schmidt, S., Cho, S.-J., & Duff, M. C. (2023). Reduced on-line speech gesture integration during multimodal language processing in adults with moderate-severe traumatic brain injury: Evidence from eye-tracking. Manuscript submitted for publication.

Covington, N. V., & Duff, M. C. (2021). Heterogeneity is a hallmark of traumatic brain injury, not a limitation: A new perspective on study design in rehabilitation research. *American Journal of Speech-Language Pathology*, *30*, 974–985. https://doi.org/10.1044/2020AJSLP-20-00081

Cox, D. R., & Snell, E. J. (1989). *Analysis of binary data* (2nd ed.). Chapman and Hall.

Dahlberg, C., Hawley, L., Morey, C., Newman, J., Cusick, C. P., & Harrison-Felix, C. (2006). Social communication skills in persons with post-acute traumatic brain injury: Three perspectives. *Brain Injury*, *20*, 425–435. https://doi.org/10.1080/02699050600664574

Diggle, P. J., Liang, K.-Y., & Zeger, S. L. (1994). *Analysis of longitudinal data*. Oxford University Press.

Durbán, M., Harezlak, J., Wand, M. P., & Carroll, R. J. (2005). Simple fitting of subject-specific curves for longitudinal data. *Statistics in Medicine*, *24*, 1153–1167. https://doi.org/10.1002/sim.1991

Fasiolo, M., Nedellec, R., Goude, Y., & Wood, S. N. (2020). Scalable visualisation methods for modern generalized additive models. *Journal of the Royal Statistical Society (B)*, *29*, 78–86.

Fine, K. L., Suk, H. W., & Grimm, K. J. (2019). An examination of a functional mixed-effects modeling approach to the analysis of longitudinal data. *Multivariate Behavioral Research*, *54*(4), 475–491. https://doi.org/10.1080/00273171.2018.1520626

Fokianos, K., & Kedem, B. (2003). Regression theory for categorical time series. *Statistical Science*, *18*, 357–376.

Gao, X., Shahbaba, B., & Ombao, H. (2018). Modeling binary time series using gaussian processes with application to predicting sleep states. *Journal of Classification*, *35*, 549–579. https://doi.org/10.1007/s00357-018-9268-8

Green, P. J., & Silverman, B. W. (1994). *Nonparametric regression and generalized linear models: A roughness penalty approach*. Chapman & Hall. https://doi.org/10.1007/978-1-4899-4473-3

Gu, C. (2013). *Smoothing spline ANOVA models*. Springer.

Guo, W. (2002). Functional mixed effects models. *Biometrics*, *58*, 121–128. https://doi.org/10.1111/j.0006-341X.2002.00121.x

Hadar, B., Skrzypek, J. E., Wingfield, A., & Ben-David, B. M. (2016). Working memory load affects processing time in spoken word recognition: Evidence from eye movements. *Frontiers in Neuroscience*, *10*, 221. https://doi.org/10.3389/fnins.2016.00221

Heitmeier, M., Chuang, Y.-Y., & Baayen, R. H. (2023). How trial-to-trial learning shapes mappings in the mental lexicon: Modelling lexical decision with linear discriminative learning. *Cognitive Psychology*, *146*, 101598.

Hsiao, C. (2003). *Analysis of panel data* (2nd ed.). Cambridge University Press.

Huitema, B. E. (2011). *The analysis of covariance and alternatives: Statistical methods for experiments, quasiexperiments, and single-case studies*. Wiley.

Ito, A., & Knoeferle, P. (2022). Analysing data from the psycholinguistic visual-world paradigm: Comparison of different analysis methods. *Behavior Research Methods*. https://doi.org/10.3758/s13428-022-01969-3

https://doi.org/10.1007/psy.2024.27 Published online by Cambridge University Press

Kedem, B., & Fokianos, K. (2002). Regression models for binary time series. In M. Dror, P. L'Ecuyer, & F. Szidarovszky (Eds.), *Modeling uncertainty. International series in operations research & management science* (Vol. *46*, pp. 185–199). Springer. https://doi.org/10.1007/0-306-48102-29

Keele, L., & Kelly, N. (2006). Dynamic models for dynamic theories: The ins and outs of lagged dependent variables. *Political Analysis*, *14*, 186–205.

Lin, X., & Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society*, *61*(2), 381–400. https://doi.org/10.1111/1467-9868.00183

Marra, G., & Wood, S. N. (2012). Coverage properties of confidence intervals for generalized additive model components. *Scandinavian Journal of Statistics*, *39*(1), 53–74. https://doi.org/10.1111/j.1467-9469.2011.00760.x

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language*, *59*, 475–494. https://doi.org/10.1016/j.jml.2007.11.006

Oleson, J. J., Cavanaugh, J. E., McMurray, B., & Brown, G. (2017). Detecting time-specific differences between temporal nonlinear curves: Analyzing data from the visual world paradigm. *Statistical Methods in Medical Research*, *26*, 2708–2725. https://doi.org/10.1177/0962280215607411

Porretta, V., Kyröläinen, A.-J., van Rij, J., & Järvikivi, J. (2017). Visual world paradigm data: From preprocessing to nonlinear time-course analysis. In I. Czarnowski, R. Howlett, & L. Jain (Eds.), *Intelligent decision technologies 2017. IDT 2017. Smart innovation, systems and technologies*. Springer. https://doi.org/10.1007/978-3-319-59424-825

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Ram, N., & Grimm, K. (2007). Using simple and complex growth models to articulate developmental change: Matching theory to method. *International Journal of Behavioral Development*, *31*, 303–316. https://doi.org/10.1177/0165025407077751

Ramsay, J. O., & Silverman, B.W. (2002). *Applied functional data analysis: Methods and case studies*. Springer. https://doi.org/10.1007/b98886

Ramsay, J. O., & Silverman, B. W. (2005). *Functional data analysis* (2nd ed.). Springer. https://doi.org/10.1007/b98888

Sørensen, O., Fjell, A. M., & Walhovd, K. B. (2023). Longitudinalmodeling of age-dependent latent traits with generalized additive latent and mixed models. *Psychometrika*, *88*, 456–486. https://doi.org/10.1007/s11336-023-09910-z

Staicu, A.M., Islam, M. N., Dumitru, R., & van Heugten, E. (2020). Longitudinal dynamic functional regression. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *69*, 25–46. https://doi.org/10.1007/10.1111/rssc.12376

Suk, H. W., West, S. G., Fine, K. L., & Grimm, K. J. (2019). Nonlinear growth curve modeling using penalized spline models: A gentle introduction. *Psychological Methods*, *24*, 269–290. https://doi.org/10.1037/met0000193

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*, 1632–1634. https://doi.org/10.1126/science.7777863

van Rij, J., Hendriks, P., van Rijn, H., Baayen, R. H., & Wood, S. N. (2019). Analyzing the time course of pupillometric data. *Trends in Hearing*, *23*, 1–22. https://doi.org/10.1177/2331216519832483

Wang, S., Li, Z., & De Boeck, P. (2023). Evaluation of parameter estimates in logistic binary time-series modeling. Manuscript submitted for publication.

Wieling, M. (2018). 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*, *70*, 86–116. https://doi.org/10.1016/j.wocn.2018.03.002

Wood, S. N. (2013). On p-values for smooth components of an extended generalized additive model. *Biometrika*, *100*(1), 221–229. https://doi.org/10.1093/biomet/ass048

Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed.). Chapman & Hall. https://doi.org/10.1201/9781315370279

Wood, S. N. (2023). Package 'mgcv.' Retrieved from https://cran.r-project.org/web/packages/mgcv/mgcv.pdf

Wood, S. N., Pya, N., & Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, *111*(516), 1548–1563. https://doi.org/10.1080/01621459.2016.1180986

Yoon, S. O., & Brown-Schmidt, S. (2018). Influence of the historical discourse record on language processing in dialogue. *Discourse Processes*, *55*, 31–46. https://doi.org/10.1080/0163853X.2016.1193429

Zeger, S. L., & Qaqish, B. (1988). Markov regression models for time series: A quasi likelihood approach. *Biometrics*, *44*, 1019–1031.

https://doi.org/10.1007/psy.2024.27 Published online by Cambridge University Press

Appendix A

See Fig. 7.

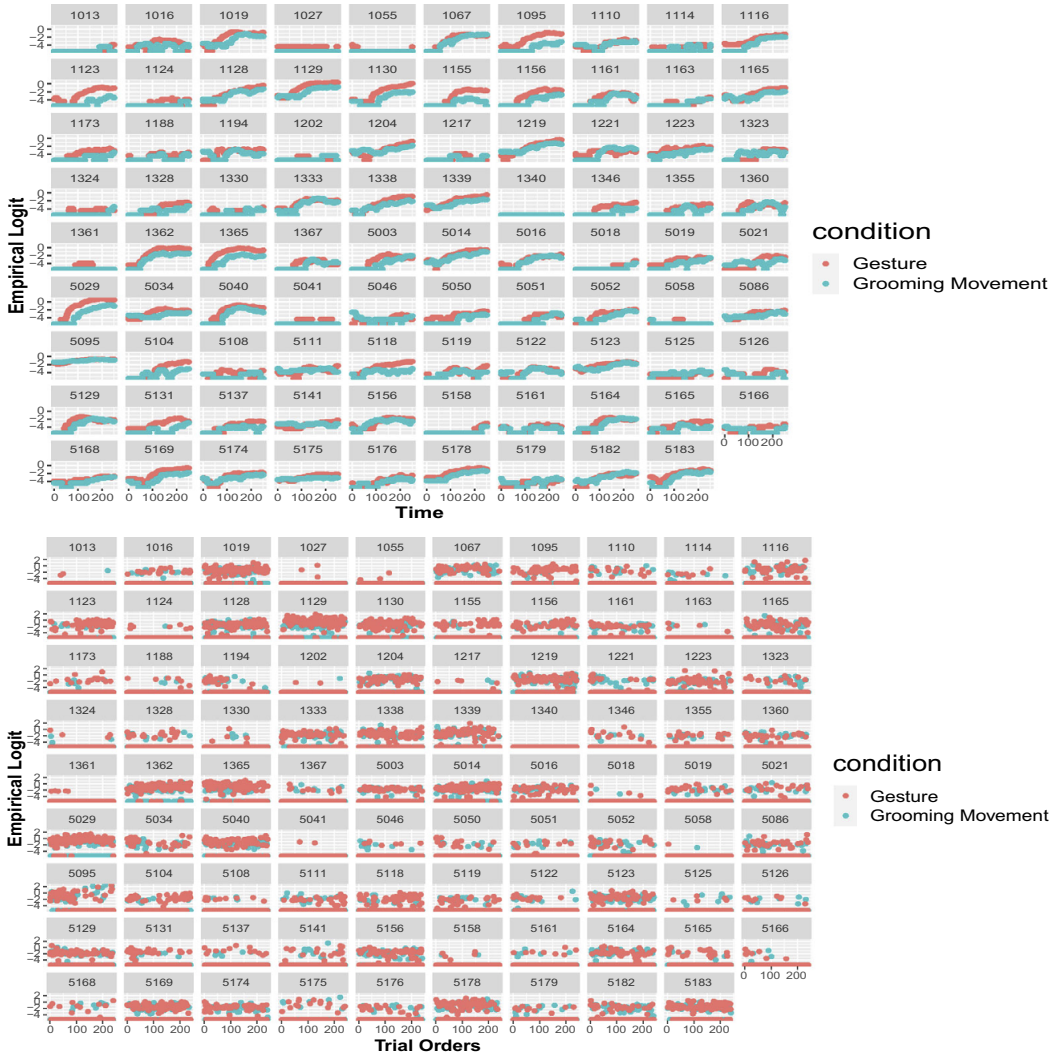

Figure 7. Empirical study: individual-level trend (top) and individual-level learning (bottom). *Note.* NC participants have an ID in the 1000's and participants with TBI have an ID in the 5000's.

https://doi.org/10.1007/psy.2024.27 Published online by Cambridge University Press

Appendix B

See Fig. 8.

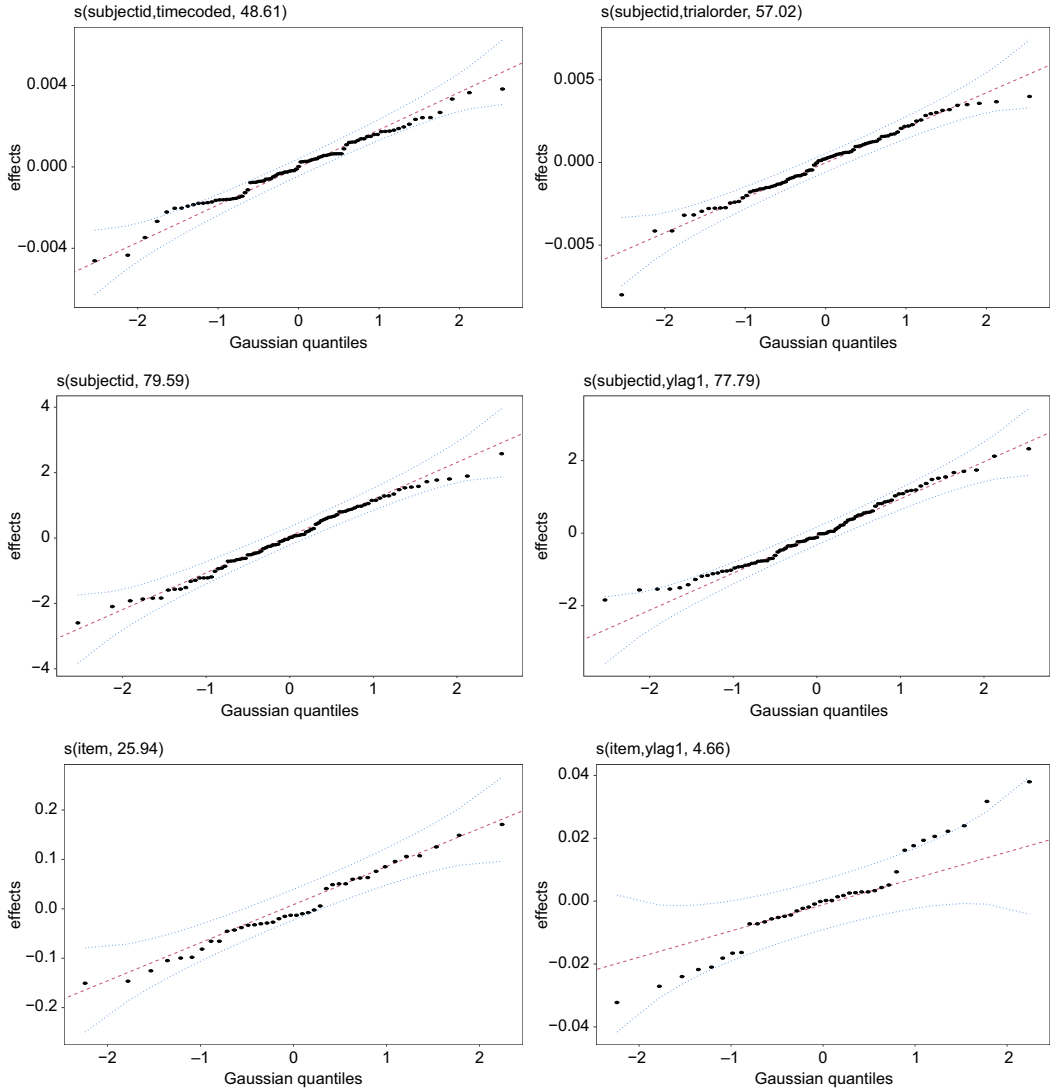

Figure 8. Empirical study: Q-Q plots of the predicted random effects in model C. *Note.* The dotted lines indicate the 95% confidence bands of the quantiles (dots).

Cite this article: Cho, S-J., Brown-Schmidt, S., Clough, S. and Duff, M. C., (2025). Comparing Functional Trend and Learning among Groups in Intensive Binary Longitudinal Eye-Tracking Data using By-Variable Smooth Functions of GAMM. *Psychometrika*, 1–30. https://doi.org/10.1007/psy.2024.27