
Book Section:
van Rijn, J., Vaci, N. orcid.org/0000-0002-8094-0902, Wurm, L.H. et al. (1 more author) (2020) Alternative quantitative methods in psycholinguistics : implications for theory and design. In: Pirrelli, V., Plag, I. and Dressler, W.U., (eds.) Word Knowledge and Word Usage: A Cross-disciplinary Guide to the Mental Lexicon. De Gruyter Mouton , pp. 83-126. ISBN 9783110517484

https://doi.org/10.1515/9783110440577-003

Jacolien van Rij, Nemanja Vaci, Lee H. Wurm and
Laurie Beth Feldman

Alternative quantitative methods in psycholinguistics: Implications for theory and design

Abstract: We describe three different methods that are appropriate to analyze various types of psycholinguistic data. We discuss some of the strengths and weaknesses of each and their suitability according to characteristics of the data. Methods include analysis of variance (ANOVA), linear mixed-effects modeling (LME) and generalized additive mixed models (GAMM).

Keywords: analysis of variance, ANOVA, linear mixed-effects modeling, LME, generalized additive modeling, GAM, model criticism, collinearity, autocorrelation, experimental design, time course data, mouse tracking

1 Introduction

Statistical analyses are an important tool for interpreting experimental results and generalizing the findings. As many different techniques are being used to investigate the structure and processing of language, there is a large variation in the types of psycholinguistic data that are being generated: for example, grammatical judgements, reaction times, ERP responses, eye gaze fixation durations, and corpus counts. These different types of data impose different constraints on the statistical methods, and consequently one style of statistical analysis is not appropriate for all types of experimental data. To facilitate choosing the appropriate statistical method, this chapter provides an overview of the regression methods that are currently used in psycholinguistics.

Jacolien van Rij, University of Groningen, Department of Artificial Intelligence, Groningen, The Netherlands
Nemanja Vaci, University of Oxford, Department of Psychiatry, Oxford, UK
Lee H. Wurm, Wayne State University, Department of Psychology, Detroit, USA
Laurie Beth Feldman, University at Albany, Department of Psychology, Albany, USA

https://doi.org/10.1515/9783110440577-003

1.1 Focus of the chapter

The "preferred" statistical method is largely determined by the nature of the data, the structure of the data, and the design of the experiment. Relevant factors are whether they are continuous data, such as reaction times, ERP responses, or fixation durations, or categorical data, such as accuracy data (i.e. binary data), type of morphological construction, or eye gaze fixation area. In this chapter we additionally make a broad distinction between behavioral data and time course data: *Behavioral data* are characterized by a single measure per trial, such as responses, accuracy, or reaction times. *Time course data*, on the other hand, consist of multiple measures per trial, which are ordered in time. Examples are EEG recordings measured while processing a word, eye gaze position during listening to a sentence, pupil size during the trial, or tongue position while producing a word. In practice, time course data are often analyzed as behavioral data by summarizing the measurements in a trial or in a certain time window to arrive at a mean value, but ideally one would like to investigate the changes over time along multiple dimensions of information. The statistical method is also determined by the design of the experiment: in a typical design (factorial experiment investigating the main effects and interactions between manipulations with only a few – often two or three – possible values) all our predictors are categorical, whereas when analyzing natural language we would like to include continuous predictors (henceforth *covariates*; numeric predictors, with an infinite range of potential values). Additionally, we may want to account for structure in the data that we are not interested in. For example, in most experimental studies the participants produce multiple responses. In such data, we would like to account for the variability introduced by the various participants, while our results should generalize over these particular participants and should provide information about the population.

This chapter focuses on the regression methods, and specifically presents linear mixed-effects modeling (LME; e.g. Pinheiro and Bates 2000; Baayen, Davidson, and Bates 2008) and generalized additive mixed modeling (GAMM; Lin and Zhang 1999; Wood 2011, 2017) as two complementary methods for analyzing most types of psycholinguistic data. LME is particularly useful for analyzing data with categorical predictors and/or continuous predictors that are linearly related with the dependent measure. GAMMs are suited for analyzing data with continuous predictors that may show a non-linear relation with the dependent measure, in addition to optional categorical or linear continuous predictors. We will introduce LME and GAMMs using an example data set to demonstrate how these new methods allow us to go beyond the typical factorial design, so as to begin to explore language behavior more dynamically.

The statistical software R version 3.4.0 (2017-04-21) was used for the analyses, with the packages *lme4* (Bates et al. 2015) for the LME analysis, the package *mgcv* version 1.8-17 (Wood 2011, 2017) for the GAMM analysis, and the package *itsadug* version 2.3 (van Rij et al. 2017) for interpretation and visualization of the GAMM analysis. The data, analysis, and code for all the graphs are available in the online Supplementary Materials at [https://www.jacolienvanrij.com/NetWordS-SupplementaryMaterials.html], along with further reading suggestions. In this chapter our aim is to provide an overview of the different methods, without presenting the actual R code.

1.2 Experimental data used as example

The data were collected by Kit Cho, Rachel Brotman and Laurie Feldman. The experiment was designed to test the effect of different accent combinations at study and test on the spoken recognition of English words. The experiment was set up as a factorial within-subjects 2x2x2 design. In a study phase, each participant was presented with pre-recorded English words spoken with either American-English or Chinese (Mandarin) as the native language. All participants were native speakers of American English. In the test phase participants were presented with the same English words, and they had to judge whether those words were produced by the same speaker as in the study phase, or by the other speaker. Thus, two manipulations were introduced in the experiment: *Accent* (accent at test phase: English or Chinese), and *Congruency* (whether or not the accent in the study and test phases matched or mismatched).[1]

Participants indicated whether the speaker was the same or different than in the study phase by moving the mouse and clicking on one of the corresponding words that was presented in the top-right or the top-left corner of the screen. The positions of the words "SAME" and "DIFFERENT" were balanced across participants. The accuracy and the reaction time were recorded, along with the mouse trajectory from the resting position (bottom-center of screen) to the appropriate word. The auditorily presented words were balanced for length (between 3 and 8 characters), with the frequency ranging between 0.044 and 325 per million words (based on the English OpenSubtitle corpus).

1 The experiment contained another manipulation: in the study phase participants had to either listen to the words, or they had to listen and repeat the words. As the effect of study task was very subtle, we ignore this manipulation for the current presentation purposes.

These example data contain the typical psycholinguistic behavioral measures accuracy and reaction time, but also the mouse trajectory (x and y coordinates over time). This time course measure has properties in common with increasingly popular online measures such as EEG, eye tracking, pupil dilation, pitch contours, and articulography. Time course measures potentially provide more information about the processing of the stimuli, but we demonstrate below that without new analytical methods much of that information is lost.

1.3 Outline

In the following sections we will show how we could analyze the responses from the mouse tracking task using traditional ANOVA (analysis of variance) and provide an overview of the more recent methods LME and GAMM in Sections 3 and 4, by using data from the same experiment. On the basis of these analyses, we will provide guidelines on when and how to use these methods. In the final sections of the chapter we will argue that one needs to be extremely careful in the interpretation of statistical results, because each of the currently available analytical methods has severe limitations. In the discussion, we delineate the implications of the statistical methods that we use, the limitations for interpretation and consequences for design.

As all methods discussed in this chapter are basically regression analyses, we will first provide a short introduction to regression analysis and list the assumptions that hold for all regression analyses.

1.4 Basics of regression modeling

Linear regression uses a linear functional relation to describe how a numerical dependent variable varies with the values of predictors. As an example, we could use a simple linear regression model to investigate the effect of *Congruency* (match or mismatch item) on response time:

$$(1) \quad y \sim \beta_0 + \beta_1 x + \varepsilon$$

The regression model describes the relation (indicated with the symbol '$\sim$') between reaction times (y, the dependent variable that is on the left-side of the '$\sim$') and the predictor Congruency (x, which is on the right-side of the '$\sim$') as a single regression line. The symbol 'ε' represents random noise, deviations from the regression line that are not fitted by the model. The line is

characterized by two parameters: β_0, a constant value called the intercept, and β_1, the slope for the effect of Congruency. The intercept specifies the height of the line, because it is the value of the dependent variable *y* when the predictor *x* equals the value 0. The slope specifies the direction of the line: it is the increase in *y* when *x* increases by 1 unit. If *x* is a categorical predictor such as Congruency ("match" and "mismatch"), each of the levels is assigned a value: "match" is represented by the value 0, and "mismatch" by 1. As a result, the slope coefficient β_1 actually models the *difference* between the reference level "match" and the level "mismatch", as illustrated in Figure 1.

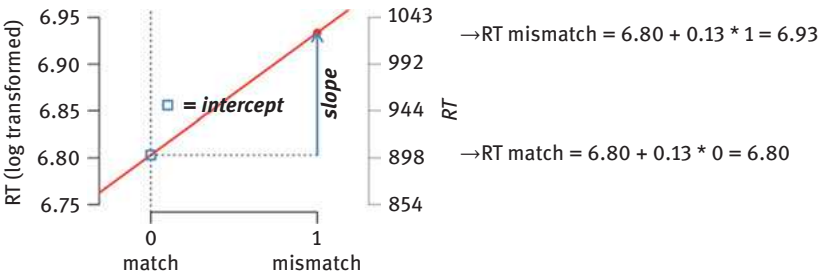

Figure 1: Schematic overview of the interpretation of linear regression coefficients.

Three assumptions for regression analysis:

i. The observations should be independent.
ii. The residuals should follow a normal distribution.
iii. The variances should be equal (often called *homoscedasticity*), which implies that the variances should be independent of the means.

The first assumption, i.e., independent observations, is violated if we do not take into account in our analysis that the data are produced by sets of participants and items. Particular participants or particular items may introduce consistent variation in the data, for example consistently slower response times than average. The assumption is also violated when we do not take into account in our analysis that the data within a time series trial are correlated. For example, in mouse tracking data the position of the mouse at the next timestamp is largely dependent on where the mouse is in the present moment. The second assumption states that the residuals should be normally distributed. The residuals are the difference between the observed data and the fitted values of the regression model. In other words, the residuals are that part of the data that is not explained by the statistical model. The third assumption specifies that the variance does

not increase or decrease with an increasing mean. This is often tested by plotting the residuals of the regression model against the fitted values. For example, the assumption of homoscedasticity is violated when the residuals show a wider spread for higher fitted values than for lower fitted values.

A typical property of experimental data collected to study language processing is that the data are structured by participants and items. Participants and items are considered as random samples from the population of speakers and from the population of words in the language. Stated differently, our focus is less on the performance of specific participants, and more on the possibility to make generalization to the whole population. Participants and items may introduce greater variation to the data than do the experimental manipulations of interest. For more precise statistical estimations, the statistical tests used in language processing ideally take into account the variation due to participants and items, so that the experimental effects are not masked by variation in participants and items.

The methods discussed in this chapter, namely repeated-measures ANOVA, linear mixed-effects modeling (LME) and generalized additive mixed modeling (GAMM) are variants of the regression model, but take into account the variability in participants and items. We refer the reader to other textbooks (e.g., Baayen 2008; Gelman and Hill 2007) for a more extensive introduction to linear regression.

2 Traditional methods in psycholinguistic research: ANOVA

This section analyzes the mouse tracking responses and reaction times using repeated-measures ANOVA, which is still one of the most frequently applied analyses in psycholinguistic research. ANOVA (acronym for analysis of variance) is particularly suited for analyzing factorial designs. The section ends with a discussion of how the use of ANOVA has shaped our experimental designs.

2.1 Introduction to ANOVA

An ANOVA tests whether the means of different groups are the same by comparing the variance *between* the groups with the variance *within* the groups using an F-test. The F-test compares the ratio of variances to the F-distribution, while taking into account the number of observations and the number of groups, to test if

the groups differ significantly from each other. One could view an ANOVA as a special case of a linear regression analysis with only categorical predictors.

As ANOVA compares group data, it is better suited for analyzing behavioral data than for analyzing on-line time course measures. For example, we could use ANOVA to analyze the accuracy and reaction times of the responses on the mouse tracking task, in order to determine whether the experimental manipulations of study-test Congruency and Accent influence the accuracy and reaction time of the response. Note that in this experiment item order is randomized and location of the match and mismatch box is counterbalanced across participants.

To account for the fact that the responses are not independent and that subsets of the data are produced by different participants and different stimuli, we use a *repeated-measures ANOVA*, which partitions out the variability due to individual differences. The input for a repeated-measures ANOVA is the means for each condition *per participant*. This is generally referred to as an F1 analysis (cf. Clark 1973). To account for the variation in items, an additional repeated-measures ANOVA on the averages *per item* (collapsed over participants) is generally performed. This is referred to as the F2 analysis. The Supplementary Materials provide more details and the code for running the analyses; here we only present the results.

2.2 In practice: Analyzing responses using RM-ANOVA

We analyze the behavioral responses of the mouse tracking data, i.e. accuracy and reaction times, using ANOVA as implemented in the R package *ez* (Lawrence 2016). For visualizing the accuracy data (Figure 2, left) proportions of correct responses were calculated. However, the underlying distribution for accuracy data is *binomial*: the accuracy of a response is correct or incorrect, or has the value 0 or 1. For analyzing binomial data, the logit transformation[2] is preferred over proportions, because ANOVA assumes normally distributed data. The proportion scale has a finite range between 0 and 1, whereas the logit scale is continuous. We included the categorical predictors *Accent*, the accent of the speaker during the test (English or Chinese), and *Congruency*, whether the word was produced by the same speaker in the training phase. *Accent* and *Congruency* are tested within participants.

2 The logit transformation is: $\text{logit} = \ln((n_{\text{correct}} + c) / (n_{\text{incorrect}} + c))$, in which $n_{correct}$ and $n_{incorrect}$ are the numbers of correct and incorrect responses and c is an arbitrary constant to avoid undefined numbers when zero counts occur (set to 0.5 here).

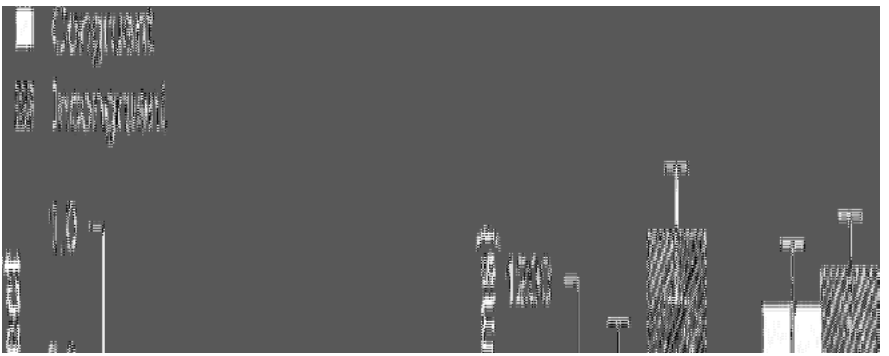

Figure 2: Accuracy (*Left*) and reaction times (*Right*) of the correct responses in the mouse tracking task. The solid bars represent the trials for which the accent in the test phase matched the accent in the study phase, the dashed bars represent the trials for which the accent in the test phase did not match the accent in the study phase. Error bars: ±1SE (i.e. standard error of the participant means).

Accuracy. The F1 ANOVA of the accuracy data indicates a significant interaction of Accent x Congruency ($F(1,32) = 4.71$, $p = .038$), and significant main effects of Congruency ($F(1,32) = 16.02$, $p = .00035$) and of Accent ($F(1,32) = 4.34$, $p = .045$). We speak of an *interaction* when the relation between a predictor and the dependent variable is changed by the value of another predictor. In this example, the effect of Accent changes depending on Congruency of speaker at training and at test. The F2 ANOVA suggests the same significant interaction: Accent x Congruency ($F(1,62) = 52.54$, $p < .001$), along with a significant main effect of Congruency ($F(1,62) = 55.12$, $p < .001$), and a marginal effect of Accent ($F(1,62) = 3.61$, $p = .062$). The effects that are significant in both the F1 and F2 analyses will be labeled as significant, which is the interaction between Accent and Congruency and the main effect of Congruency.

Reaction times. We only included correct answers in the analysis, and the reaction times were log-transformed. Reaction time data are generally not normally distributed, but rather skewed. Therefore, they are commonly transformed by taking the log, inverse, or power transformation. In contrast with the accuracy data, the F1 and F2 ANOVA analyses for the log-transformed reaction times only revealed a significant main effect of Congruency ($F1(1,30) = 8.38$, $p = .007$; $F2$ $(1,61) = 9.81$, $p = 0.003$); see Figure 2, right.

A disadvantage of ANOVA tables is that they only indicate which predictors are influencing the model estimations significantly. To interpret the direction of the interaction we could look at the accuracy plot (Figure 2, Left panel). The plot reveals that the effect of Congruency is different for the two levels of Accent: The

Chinese accented speech at test seems to result in a significant difference between match and mismatch items, but this difference seems to be absent for items pronounced with an English accent. Thus, with an unfamiliar accent, participants resort to a strategy of selecting "SAME", but with a familiar accent they do not use such a strategy.

2.3 Discussion

The repeated-measures ANOVA provides a relatively simple and quick test to confirm which factorial predictors contribute significantly to the values of the response variables. The results are easy to report, following the standard conventions in the literature. However, a disadvantage of ANOVA tables is that they only indicate which predictors are influencing the model estimations significantly. Post-hoc tests are necessary to interpret the direction of the interaction, because coefficients of the estimated effects are not automatically given. In the accuracy plot (Figure 2, Left panel) the Accent x Congruency interaction is clearly visible. The Chinese accented speech at test seems to result in a significant difference between match and mismatch items, but this difference seems to be absent for items pronounced with an English accent.

As the ANOVA test is performed on averages, it does not provide a straightforward way to deal with missing data or unbalanced designs. This is particularly an issue for our current reaction time data, from which we excluded the incorrectly answered items. Another consequence is that participant and item variation cannot be accounted for at the same time. Instead two analyses (F1 and F2) are generally performed to account for the variation in participants and items (e.g. Clark 1973). The convention is to consider a predictor significant only when the F1 and F2 both indicate that that predictor is significant. The F1 and F2 analyses are not an ideal solution to this problem (e.g. Raaijmakers et al. 1999). Baayen (2008) has pointed out that for a design where items are nested under a condition, such as words presented in an American or a Chinese accent but not both, F1 and F2 may reveal conflicting results and may result in the incorrect (too conservative) conclusion that a predictor is not significant. One more comprehensive analysis, that can account for participants and item variation at the same time, would provide a more coherent solution.

Another important disadvantage is that ANOVA only accepts categorical predictors, which means that covariates have to be converted to be categorical. In our analysis of the behavioral responses we have only included categorical predictors, but in other analyses we may want to include continuous covariates. For example, if we would like to test whether the frequency of the

word influences the behavioral response, we need to dichotomize the frequency continuum: for example, words with a frequency lower than the median frequency are labeled as "low", the other words are labeled as "high". Rather than an arbitrary division of the frequencies into two groups, it is generally preferable to treat frequency as continuous and include it as a covariate.[3]

Although the ANOVA analysis still is the most commonly used analysis in language processing research, linear mixed-effects modeling (LME) is quickly gaining in popularity because it provides a solution for many of the disadvantages of the repeated-measures ANOVA.

3 Linear mixed-effects modeling (LME)

Linear mixed-effects modeling (LME) is a linear regression analysis that does not require group averages as input, and can handle the responses of individual trials. LME is preferred over ANOVA (i) when the data result from an unbalanced design, or contain missing observations, (ii) when the dependent variable is not normally distributed, or (iii) when continuous predictors are available. With balanced factorial designs LME has comparable power to repeated-measures ANOVA (e.g. Baayen 2008; Baayen et al. 2008; Barr et al. 2013), but we still recommend LME as it does not require separate analyses for participants and items.

3.1 Introduction to linear mixed-effects modeling

In contrast to the repeated-measures ANOVA, LME accounts for the variability among participants and for the variability among items at the same time rather than in separate analyses. In LME, a distinction is made between random effects and fixed effects. *Fixed effects* are those that are expected to hold for the entire population or expected to apply to other experimental stimuli, whereas *random effects* capture variation introduced by the particular participants and stimuli that were randomly sampled from larger populations (e.g. Pinheiro and Bates 2000; Gelman and Hill 2007; Baayen et al. 2008).

In mixed-models, i.e., models including both fixed and random effects, random effects predictors are each represented by one parameter, namely the

3 We use frequency as our example, but language skill is a measure that often gets treated dichotomously and is subject to similar limitations in the ANOVA.

standard deviation associated with the random effect. The random adjustments for each individual participant (or item) are selected such that when added to the fixed effects they provide an estimate of that participant's (or item's) performance. However, the estimates are not necessarily the same as the participants' means: they are a compromise between the mean over all participants and the participant's mean, weighted for the participant's number of observations and under the constraints that the random adjustments follow a normal distribution with a mean of zero and the estimated standard deviation for the random effect (Gelman and Hill 2007). If the participant contributed only a few observations and much of the data were missing, the estimated mean for that participant will be closer to the mean of all participants than to his or her observed mean, i.e. the random adjustment for that participant will be smaller than expected. The assumption that random effects follow a normal distribution allows for making generalizations: an extremely fast reaction time, much faster than average, is atypical and is not very likely to be observed in a follow-up experiment with different participants. So, the estimated mean for such a fast participant also tends to be closer to the mean of all participants than to the observed mean for that participant. The effect that the estimations for extreme participants are closer to the overall mean than to their observed means is called *shrinkage*.

Two types of random effects can be specified in LME: random adjustments of the intercept, and random adjustments of slopes. Random intercepts adjust the height of regression lines for each participant or item. Random slopes adjust the slope of a regression line for each participant or item. Figure 3 illustrates the regression line in our earlier example, in which we used linear regression to analyze the effect of Congruency on the log transformed reaction times. In the left panel random intercepts are illustrated: the intercept adjustments raise or lower the regression line (black solid line), but do not change the relation between the congruency conditions. In the center panel random slopes are illustrated: the slope adjustments tilt the regression line, in order to change the difference between the two Congruency conditions, but does not change the height of the regression line. The right panel illustrates a combination of random intercepts and random slopes. For one of the participants an increase in intercept but a decrease in slope is estimated (higher gray dashed line). This participant is slower in responding, but does not show much difference in response times between the match and mismatch trials. The lower dashed line represents a faster participant, but with a stronger effect of Congruency: the intercept is much lower than the average intercept, but the slope is increased. In short, random intercepts capture general differences in performance between participants (or items) and random slopes capture variation between conditions for those participants (or items).

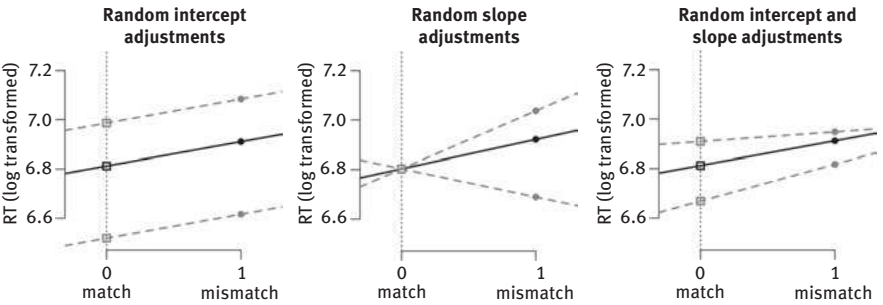

Figure 3: Schematic overview of random intercepts (*left panel*), random slopes (*center panel*), and the combination of random intercepts and slopes (*right panel*).

Different from an ANOVA analysis, LME does not return which predictors or random effects are significantly contributing to the model. Rather, given that these predictors are included, the outcome indicates whether or not the contrasts are different from the intercept and whether or not the slopes are significantly different from zero. A model comparison procedure is necessary to determine which predictors significantly contribute to the model. The collective wisdom is to start by determining the appropriate random effects structure, and then to test which fixed effects are significant. Backward-fitting model comparison procedures, which start with the most complex fixed-effect model and gradually reduce non-significant interactions and predictors, generally reduce the risk of overlooking interactions and main effects (e.g. Barr et al. 2013). R packages are available that facilitate model comparison procedures by automatic model selection.

3.2 In practice: Analyzing responses using LME

Here, we re-analyze the behavioral responses of the mouse tracking data, i.e. accuracy and reaction times, using LME as implemented in the R package *lme4* (Bate et al. 2015), and the R package *multcomp* (Hothorn, Bretz and Westfall 2008) for inspection of the model estimates.

Accuracy. When the raw data rather than averages are being analyzed, LME does not require transformation of binomial data. LME implements generalized algorithms to analyze binomial data, or data from several other non-Gaussian distributions. In our example, backward-fitting model comparison procedures were used to determine the maximum random effects structure that was allowed by the data, and to determine which fixed effects to include. The

random effect structure for the accuracy data included random by-participant slopes for Accent and Congruency and the interaction Accent x Congruency. The best-fitting model included a significant interaction between Accent and Congruency ($\chi^2(1) = 4.77$, $p = 0.029$). The model coefficients explain this interaction: Participants do not differ from chance performance on items pronounced with an English accent ($\beta_{intercept} = 0.118$, $SE = 0.214$, z-value $= 0.550$, $p > 0.1$), and do not show a difference between congruent and incongruent items with an English accent ($\beta_{Mismatch} = -0.081$, $SE = 0.394$, z-value $= -0.204$, $p > 0.1$). However, they do show a significant difference between congruent and incongruent items with an Chinese accent ($\beta_{Mismatch:Chinese} = -1.397$, $SE = 0.621$, z-value $= -2.250$, $p = 0.024$): Participant's performance on congruent items with a Chinese accent is significantly more accurate than English congruent items ($\beta_{Chinese} = 0.920$, $SE = 0.353$, z-value $= 2.608$, $p = 0.009$), but the performance on incongruent items with an Chinese accent is not significantly different from the performance on incongruent items with English accent ($\beta_{MismatchCH\text{-}MismatchEN} = -0.477$, $SE = 0.301$, z-value $= -1.586$, $p > .1$).

However, the models with this random effects structure were showing difficulties to converge. Therefore, we replaced the random effects by a random intercept adjustment for each combination of Congruency, Accent, and Participant to reduce the number of variance and correlation parameters. This alternative random effects structure yielded the same conclusions.

Reaction times. We analyzed the reaction times of the correct responses. The reaction times are log transformed to improve normality. A backward-fitting model comparison procedure suggested inclusion of a random intercept for participant, and by-participant random slopes for Congruency and Trial (centered and scaled, to facilitate the interpretation of the regression coefficients). The slope for Trial was included to account for correlations between subsequent reaction times (e.g. Baayen and Milin 2010). Only the main effect of Congruency was found to be significant ($\chi^2(1) = 7.850$, $p = 0.005$): the correctly answered incongruent items are responded slower than the correctly answered congruent items ($\beta_{Mismatch} = 0.087$, $SE = 0.030$, t-value $= 2.90$).

3.3 Discussion

Basically, the LME analyses of behavioral responses lead to the same results as the repeated-measures ANOVA. A large advantage of LME over repeated-measures ANOVA, however, is that it combines the participants and item analysis into a single statistical model. Further advantages are that LME does not require separate

post-hoc tests, as the coefficients of the estimated values are provided in the summary, and that covariates can be included in the analysis.

A disadvantage of LME is that one needs to determine the appropriate random effects structure. Only including random intercepts for participants and items without an adjustment for the experimental conditions may result in over-confident estimates of the fixed effects, finding effects that are not really there (e.g. Pinheiro and Bates 2000; Baayen, Davidson and Bates 2008; Barr et al., 2013). For example, it is not uncommon to add measures of vocabulary and spelling knowledge. However, these should be contrasted with simple random slopes and intercepts to make sure that the estimated effects are not caused by random variation between participants and items. To avoid over-confident estimates that are not generalizable, Barr et al. (2013) argue to maximize the random effect structure based on the experimental design. This means including the slopes for all experimental predictors by participants and items in addition to the random intercepts. Recently, Bates et al. (2015) showed that this is in practice not possible for many data sets. Missing data and limited data samples strongly limit the number of random effects that lead to a reliable and converging estimation of the parameters of the model (see also Baayen et al. 2017 and Matuschek et al. 2017). Determining the appropriate random effects structure is one of the challenges when using LME.

As LME can also include covariates, it seems at first glance to be the obvious choice for analyzing the mouse tracking data. LME even allows polynomial functions (or other non-linear functions) for modeling a non-linear relationship between the dependent variable and a covariate (see Supplementary Materials). However, we prefer to use GAMM over LME, as will be explained in Sections 4 and 5. In the next section, we will introduce GAMMs and illustrate how they could be applied to analyze time course data, such as mouse tracking data.

4 Generalized Additive Modeling (GAM)

Language processing research increasingly makes use of time course measures to investigate *online* language processing, i.e. the actual processing of the word or utterance from the moment it is being read or heard. *Time course* measures provide multiple data samples during a trial, often with a fixed sampling rate. Traditionally, time course measures are simplified to one value per trial, for example the mean value in a specific time window or the average deviation of the mouse trajectory, to be able to perform an ANOVA. However, as noted above, this considerably reduces the information these measures provide.

Instead, we prefer to analyze the time course directly. For example, we could analyze how participants in the mouse tracking experiment move their mouse to the response location on the screen. The mouse position is a continuous measure developing over time that may reflect uncertainty and hesitation in the form of pauses and deviations from the ideal trajectory (straight path to answer location). The example data contains 101 samples per trajectory, each of which records the x-position and y-position of the mouse, and the time relative to the offset of the word. For the current data, we normalized the time between the onset of the movement and click/answer, as rate of mouse movement varies along the trajectory. The mouse movement *duration* is the time from the onset of the movement until the participant clicked to respond. Below we present an analysis of the mouse trajectory on correctly answered trials only, to facilitate interpretation. As we do not know the cause of errors, trials that are incorrectly answered were excluded from analysis. Of the 2081 trials 958 were incorrectly answered and excluded (54% of the trials were included in the analysis).

The raw data of the mouse position (x and y coordinates) are presented in the Left panel of Figure 4. Location of match and mismatch responses is counterbalanced across participants. The black dots show one sample mouse trajectory. The Right panel of Figure 4 shows a measure derived from the x and y coordinates, namely the *distance to the clicked target*, with the same sample trajectory in black. For each data point the Euclidean distance to the target, i.e. the answer that participants eventually clicked, was calculated.[4] The idea behind using the Euclidean distance is that when participants are uncertain or change their mind during response selection they take a less direct route to the target than the optimal straight path. For example, participants may initially go toward one of the responses, but, during the mouse movement, change their mind and abruptly shift to the other response, called x-flips (see example trial in Left panel of Figure 4; Freeman, Dale and Farmer 2011; Freeman and Johnson 2016). These hesitations show up in various measure as pauses or increased distance to the target (see Right panel of Figure 4).

4 The choice of distance to target as the dependent variable instead of the X and Y coordinates was made for illustration purposes, to make the analysis more comparable to other psycholinguistic time course measures such as EEG or pupillometry data.

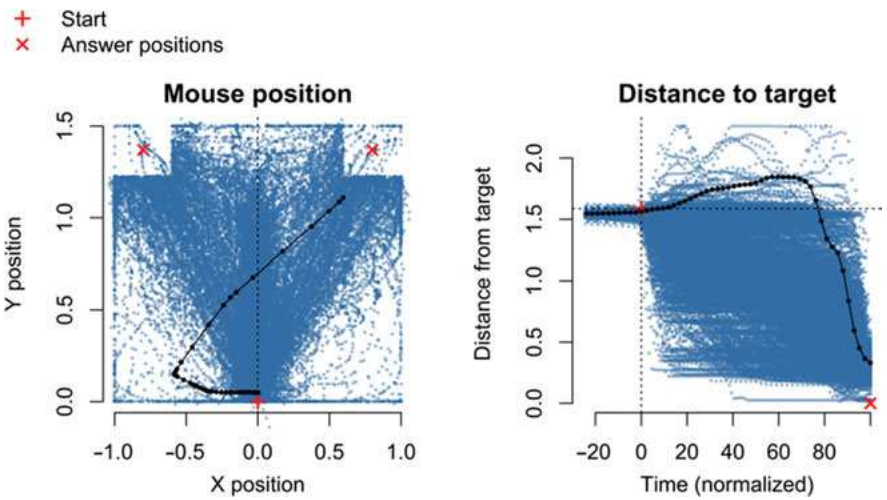

Figure 4: Mouse tracking data exclusively including correctly answered trials included. *Left panel:* Recorded X and Y position of the mouse. The black dots highlight one example trial. *Right panel:* Distance to the clicked target over time. The black dots highlight the same trial as in the left plot.

4.1 Introduction to generalized additive (mixed) modeling

Generalized Additive Mixed Modeling (GAMMs; Lin and Zhang 1999; Wood 2017) is a recently introduced analysis method that is specially designed to model non-linear covariates: it is a non-linear mixed-effects regression method, which can fit non-linear regression lines to the data. GAMMs are implemented in the R package *mgcv* (Wood 2017, 2011). In contrast with LME, the user does not need to specify the shape of the non-linear regression line (e.g., which order polynomial to use), because the model determines the non-linear pattern based on the data. The use and interpretation of GAMMs is slightly different from linear regression models. Where linear regression models aim to explain the data by fitting the *coefficients* in the regression formula, GAMMs try to optimize the smooth function that describes the potentially non-linear relation between the predictor and the dependent variable; see the formulas in Example (2).

(2) Difference between linear regression and non-linear regression (with y the dependent variable, x a predictor, β_0 the intercept, $\beta_{>0}$ the slope(s), and ε the residuals):
   - Linear regression formula: $y \sim \beta_0 + \beta_1 x + \varepsilon$

- Linear regression with *n*th order polynomial curve: $y \sim \beta_0 + \sum_{i=1}^{n} \beta_i x^i + \varepsilon$
- GAMMs: $y \sim \beta_0 + f(x) + \varepsilon$

The output of a GAMM only presents the coefficients for the linear predictors, including the intercept, i.e. the height adjustment of regression lines, intercept adjustments, and linear slopes. The output does not present a description of the *non-linear* regression lines, because the smooth functions ($f(x)$) often cannot be captured by a few coefficients. Instead the summary provides information on the wiggliness of the regression line, and whether the line is (somewhere) significantly different from zero. Visualization is necessary for interpreting the non-linear terms.

Similar to LME, in GAMMs fixed effects and random effects can be specified. However, the structure of the random effects in GAMMs is different from the random effects in LME: In addition to random intercepts and random slopes, GAMMs also provides the option to include *random smooths*, non-linear random adjustments of a regression line. These random smooths capture also random intercepts and slopes, so they are generally not combined with random intercepts and slopes for the same predictors. Figure 5 illustrates how the random effects of two different participants (Left panel) alter the non-linear fixed effect regression line to generate estimates for these two participants. It is important to realize that the random effects in the Left panel are *adjustments* of the fixed effects, with a negative value indicating a shorter distance than the general trend (represented by the fixed effect), and a positive value indicating a longer distance than the general trend.

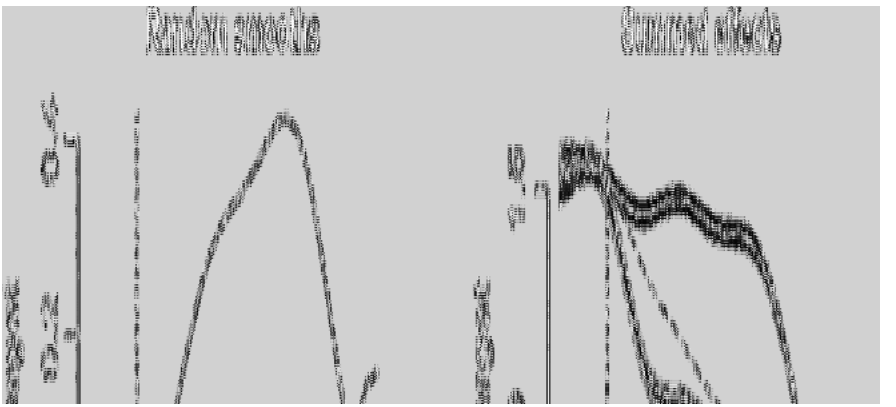

Figure 5: Non-linear random effects in GAMMs. *Left panel:* random smooths for two arbitrary participants in the mouse tracking data. Negative values indicate movement away from target. *Right panel:* summed effects for the same two participants. The random effects (Left panel plot) are added to the fixed effects smooth (dashed line in the Right panel plot) and the intercept.

Another difference with LME is the possibility of GAMMs to add *non-linear interaction surfaces*. For example, in the mouse tracking data we could include a non-linear interaction between Time, the normalized time along the trajectory (ranging between 0 and 100), and Duration, the actual duration of the trajectory in milliseconds (log transformed, ranging between 47 and 6365 ms) starting with the first mouse movement until the response. A linear interaction would imply that the slope of the regression line for Time is increased or decreased with Duration in a constant way. A non-linear interaction allows the shape of the non-linear regression line for Time to change depending on the value of Duration in a non-linear way. An example of a non-linear interaction is provided in the next section.

GAMM provides the same advantages as LME with respect to missing data and unbalanced designs. The method also includes an extensive list of link functions for handling data that is not normally distributed. Similar to LME, model comparison procedures are used to determine the best-fitting model. However, the output tables do not provide precise information on the shape of nonlinear regression lines or interaction surfaces, but visualization is necessary for interpreting the results.

In sum, advantages of GAMMs over LME are the possibilities to fit non-linear regression lines and surfaces without a priori assumptions on the shape of the regression lines. In addition, the visualization methods facilitate interpretation, whereas the polynomial terms in linear regression are rather difficult to interpret (see Supplementary Materials). Moreover, GAMMs also allow for non-linear random variations in time course patterns for individual participants and items, which result in more generalizable time course estimations (Baayen et al. 2018; van Rij et al. 2019).

4.2 In practice: Analyzing mouse tracking data using GAMMs

To investigate whether and how Accent and Congruency influenced the mouse trajectory during response selection, we analyzed the Euclidean distance to the target (see Figure 4, Right) as a dependent variable with a GAMM analysis. To account for differences in strategy by participants and conditions, which are reflected in the paths to the target, we included the time course along the trajectory per participant per condition (predictor Time) as a non-linear random smooth. Similarly, the time course per word per condition was included as a non-linear random effect to account for differences in processing of different words. In addition, we included a random intercept adjustment per *event*, i.e., a unique participant-trial combination. Individual trials are likely to show

variation in time series data, as each trial consists of multiple measurements. Including an intercept for each event accounts for the variation between trials.

After determining the random effects structure, a backward-fitting model comparison procedure was used to determine the effect of Accent and Congruency. Model comparison procedures for GAMMs are less easy to interpret than for LME, as the models that differ minimally are not necessarily strictly nested. When doing model comparisons with linear regression, we try to compare two models that differ in only one term: one of the two models contains an additional term, which the other model lacks. These models are called *nested*. However, with GAMMs the difference of one model term does not necessarily mean that the two models are nested, because the shape of the smooth terms may change non-linearly (for example by changing the number of base functions being used) in the presence or absence of other model terms. In other words, the model with fewer model terms does not necessarily end up being the simplest model. Therefore, visualization and checking the summary output provide useful information in addition to the model comparison results themselves.

Visual inspection. We start with a model that includes a three-way interaction between Accent, Congruency, and *Time* along the trajectory (order of mouse positions, with values between 0, indicating the start of the movement, and 1, the response click). As Time is the only continuous predictor of these three, the interaction was implemented as a non-linear regression line for Time split by a four-level grouping predictor representing the two-way interaction between Accent and Congruency. Beside this non-linear interaction, we included the *Duration* (log transformed; the total time duration of the mouse movement until the response click) of the trajectory as a non-linear main effect and the (additive) non-linear interaction between Time and Duration. By normalizing the mouse trajectories, the differences between fast and slow trials are lost. The interaction between Time and Duration captures potential spatial differences in trajectory that are related to the duration of the trajectory (paths tend to be straighter when velocity is high).

Figure 6 plots the estimated effects for Chinese accented words (Left panel) and English accented words (Center panel). The straight solid line indicates a straight ideal path to the target. From timestamp 40 (around 40% of the trajectory) the average mouse trajectories in all conditions deviate significantly from a straight line. The Right panel of Figure 6 plots the estimated differences between the accents for match items (solid line) and mismatch items (dashed line) with 95% confidence interval. A positive difference indicates that the mismatch items deviate more from the ideal trajectory than the match items. A negative difference indicates that the match items deviate more from the ideal

trajectory than the mismatch items. Although the difference lines deviate from zero in the second half of the trajectory, the difference between the accents does not become significant as the zero line is always included within the confidence bands. The differences between the Congruency conditions within Accent (not visualized here) are also not significant based on the visualization of the model's estimates.

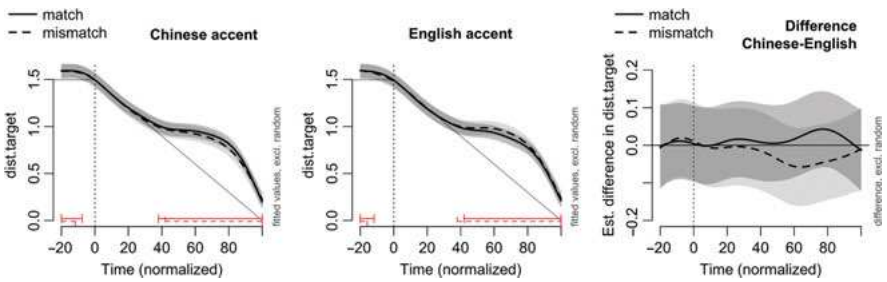

Figure 6: Estimated effects for Chinese accented words (*Left panel*) and English accented words (*Center panel*). The straight diagonal solid line indicates an ideal (straight) path to the target. The red horizontal interval markers indicate at which parts of the trajectory participants significantly deviate from this straight trajectory. The *Right panel* shows the differences between Chinese and English items for the match (solid line) and mismatch items (dashed line) with 95% confidence interval visualized by shading.

Model comparison. Visualization is an important tool for significance testing with GAMMs. Another important tool is a model comparison procedure. Here, we compared the model with the effects of Congruency and Accent (four-level categorical predictor[5]) with a model that does not include these effects using a Chi square test on the fREML scores, i.e. the minimized smoothing parameter selection score, while taking into account the difference in degrees of freedom specified in the model. The model without the effects of Congruency and Accent is preferred, because it has a lower fREML score (difference of 36.863) and lower degrees of freedom (6 df), supporting the earlier conclusion that there is no difference in trajectory for the Congruency and Accent conditions. Note, however, that fREML scores (default selection score in GAMMs) are actually not ideal for comparing different fixed effects structures (see Supplementary Materials). A model comparison based on AIC (Akaike's Information Criterion) prefers the model *with* Congruency and Accent included (AIC difference of 3.10). Sections 4 and 5 explain why

5 We also tested breaking apart the four-level predictor into separate two-level predictors Congruency and Accent, which resulted in the same conclusions. See Supplementary Materials.

different significance tests can point to opposite conclusions and we provide suggestions on how to deal with a situation of inconsistent information.

Non-linear interactions. Besides the effects of Congruency and Accent over Time, we also included a non-linear interaction between Duration and Time, for which the estimated effects are illustrated in Figure 7. The contour plot (Left panel) can be read like a hiking map with the contour lines and the colors indicating the height: blue areas are valleys and the yellow areas hills. The right panel shows the estimated regression lines for mouse trajectories with durations of 5 and 7 (log scale). The two plots suggest that participants use different strategies in short (e.g., Duration of 5) and long trajectories (e.g., Duration of 7.5). The long trajectories move with a direct path (indicated by straight diagonal line) towards the target until half-way, and only then seem to reconsider their choice, i.e. the position does not decrease for quite a while. Short trajectories follow a less straight path initially, but do not seem to hesitate half-way through. Contour plots are a useful instrument to interpret non-linear interactions. Although higher order non-linear interactions (3-way or higher) are possible in GAMMs, they get increasingly more difficult to visualize and interpret.

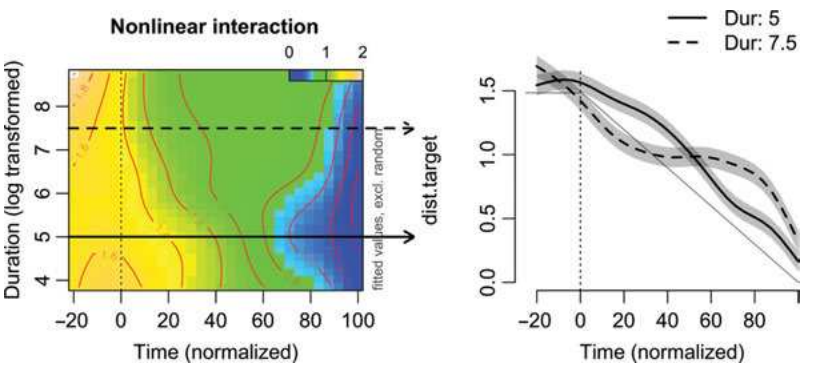

Figure 7: *Left panel:* Contour plot visualizing a non-linear interaction between two continuous predictors (Time and Duration). The colors and contour lines indicate the distance to the target. *Right panel:* estimated distance to target for Durations of 5 (148 MS) and 7.5 (1808 MS) over Time. The straight line indicates the ideal path to the target.

4.3 Discussion

To summarize, GAMMs are particularly suited to analyze non-linear patterns and time course data, because they allow us to fit non-linear regression lines, non-linear interactions, and non-linear random effects. As the non-linear

effects are not represented with coefficients, the statistical method necessarily relies on visual inspection of the model estimates, which facilitates interpretation and increases understanding of patterns in the data in comparison with linear regression analyses.

An important contribution of GAMMs for the analysis of time course data is the possibility to investigate different questions, such as investigating at which moment the trajectories of different conditions start to differ, or when trajectories start to deviate from the ideal path to the target. For mouse tracking data, these questions are currently investigated with the calculation of a separate t-test for every time bin or area under the curve between ideal trajectory and observed one (e.g. Freeman and Ambady 2010). Disadvantages of GAMMs are that different sources of information, such as visualization and model comparisons, need to be assessed to determine whether a predictor contributes significantly to the model; that models may take a long time to run; and that the estimated effects often cannot simply be described with a single coefficient.

5 All statistical models are wrong

In the previous sections we have provided an overview of different regression methods for language processing research. The traditional repeated-measures ANOVA is a powerful analysis for behavioral data of factorial experiments with balanced designs and no missing data. However, with unbalanced or nested designs, missing data, continuous covariates, or not normally distributed dependent variables, a mixed-modeling approach is a better choice. Linear mixed-modeling has the advantage of returning interpretable coefficients with their statistics which make it easier to quickly quantify linear effects. For time series data and data with non-linear trends, generalized additive mixed modeling provides more explanatory power and the most precise data fit.

However, no statistical model is perfect. Problems with statistical models are generally detected when evaluating the model. Therefore, *model criticism* is the most important part of statistical analyses. This involves inspection of the residuals and testing the generalizability of the model. The first thing to check is the assumptions of regression models: (i) are the residuals normally distributed? (ii) and are the observations independent? We have already listed disadvantages for all the discussed regression methods, but in the next sections we explain some more fundamental problems with regression models and how they influence the reliability of the analyses.

5.1 Power of the model

In this chapter we have presented the GAMM analysis of the mouse tracking data, on the basis of which we concluded that the mouse tracking trajectories (or rather the Euclidean distance to the target) do not differ significantly between the words with English and Chinese (Accent), nor between the items that were congruent and incongruent in accent in comparison with the study phase (Congruency). However, the absence of a significant effect for Accent and Congruency could have various causes, such as participants' mouse trajectories are really not influenced by Accent or Congruency, or there is not sufficient power to detect the effect, or the model is not a good fit of the data and fails to include important structure in the data. In some cases, some statistical methods may come to another conclusion.

For example, we could run a LME model with Time modeled as a *non-linear* polynomial effect (cf. Growth Curve Analysis; e.g. Mirman, Dixon and Magnuson 2008; Mirman 2014) as an alternative to the GAMM analysis. To fit the non-linear trend of Time, we include a fourth order polynomial. This means that the Euclidian distances over Time are fitted with a quadratic function. This LME model includes the Time variable raised to the power of 1, 2, 3 and 4, and in addition Congruency (match or mismatch), Accent at test (English or Mandarin), and Duration (log transformed duration of the mouse movement) as fixed effects predictors. The complete analysis is part of the Supplementary Materials. In contrast with the GAMM analysis, the LME model with a polynomial effect for Time indicates that the mouse trajectories of the different conditions do vary significantly. The polynomial LME model suggests that the words with an English accent elicit *more* uncertainty with respect to whether or not the accents in the study and test phase match, and also produce a less direct path to the answer compared with Chinese accent, whereas in the GAMM the interaction between Congruency, Accent, and Time does not reach significance. How do we know whether this effect is just an artifact of the analysis or it exists in reality? Stated bluntly, this effect could be a false positive finding where we are wrongly concluding that there is an effect, when there is none (Type I error).

In this case it may be more constructive to ask first the opposite question: assuming that the effect exists in the population, how likely is it for us to detect the effect in the sample and to observe a statistical difference between the trajectories? The statistical procedures that we introduced throughout the chapter differ in how powerful they are to deal with particular types of data. To investigate this question further, we simulated hypothetical trajectories that are similar to the collected data (see Supplementary Materials).

Simulations. The basic shape of the distance function across time was simulated with a logistic curve for every participant, thus, the simulated data observe a sigmoidal shape. More importantly, the intercept of the function differs between subjects. In the next step, we added more noise to the data, but also a categorical predictor that has a small interaction with time course of the experiment, shifting the trajectory of one condition up (0.05 for simulated Euclidean distance over time). Our simulated population comprised 300 subjects, with a two level factor (match or mismatch) and 121 time points for each condition. Five different models were estimated on every subset of the population, starting with a subsample of only two subjects in the analyses and increasing up to the moment when the whole population was sampled. We used the following models:

i. LM Linear: a linear regression model with a linear effect for Time.
ii. LM Polynomial: a linear regression model with polynomial effects for Time.
iii. LME Linear: a LME with a linear effect for Time and a by-subject intercept adjustments.
iv. LME Polynomial: a LME with polynomial effects for Time and by-subject intercept adjustments.
v. GAMM: a GAMM model with a non-linear effect of time and by-subject intercept adjustments.

Finally, for every iteration, that is a subset of the population, we monitored outcomes of the models for 100 separate simulations. These outcomes were used to calculate the proportion of the obtained significant effects, thus, its power.

The results are illustrated in Figure 8. The simulations indicate that the regression models with a linear effect for Time (LM Linear and LME Linear) require vastly more subjects to be powerful enough, that is, to detect the effect in 80% of the simulations (over 300 subjects). The polynomial models in the case of simple regression (LM Polynomial) need to sample relatively fewer subjects, approximately 60 of them. Thus, specifying polynomial effects in the model explains additional variance, making the model more powerful. The most powerful in estimating the simulated interaction are the linear mixed-effects modeling with polynomial effects (LME Polynomial) and GAMMs. They need approximately 55 subjects to have 80% power for the effect estimation. To summarize, these simulations show that for detecting this simulated interaction a non-linear regression line is crucial.

Model criticism. Inspection of the residuals may also reveal that a non-linear predictor should be included instead of a linear predictor. For illustration purposes we modeled one participant's mouse tracking data (Euclidean distance to target) with a GAMM and with a comparable LME model with Time (centered and scaled) included as a linear predictor (LME Linear). Figure 9 (Left

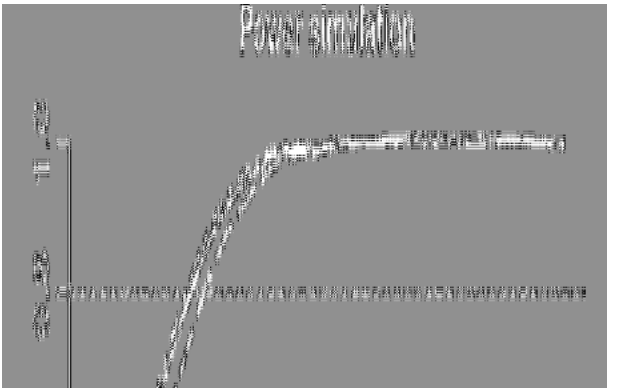

Figure 8: The power estimation for each of the illustrated analysis. X-axis represents number of subjects sampled from the population, while Y-axis represents the percentage of significant tests out of 100 simulations. The dotted horizontal line at the 0.8 value of the Y-axis indicates the moment when the statistical procedure catches the effect in 80% of the simulated times. LM – linear effect: linear regression with specified linear relation between Time course of the experiment and simulated Euclidean distances. LM – polynomial effect: linear regression with specified fourth polynomial relation. LMER – linear effect: linear mixed-effect modeling with linear effect. LMER – polynomial effect: linear mixed-effect modeling with polynomial relation. GAMM: generalized additive modeling with non-linear effect of Time.

and Center panels) plots the residuals against the values of the predictor Time for the GAMM analysis for the LME Linear analysis. Note that the residuals of the GAMM model do not show a trend over the Time values, but the residuals for the LME Linear model do. This indicates that there is unexplained structure in the residuals. The Right panel of Figure 9 shows the same plot for the GAMM model of all mouse trajectories.

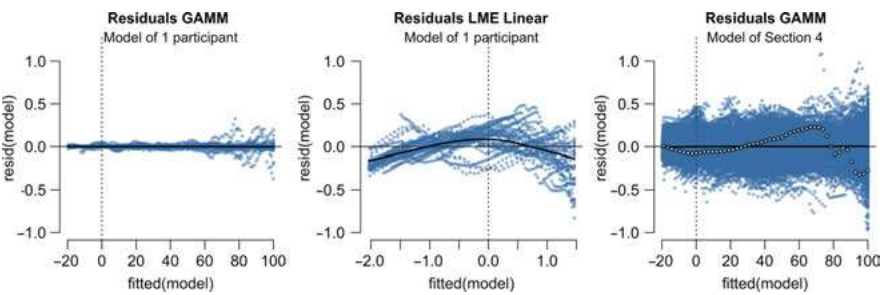

Figure 9: The residuals plotted against the predictor Time. Left: GAMM model of 1 participant. Center: LME model with linear predictors modeling the same participant. The residuals show that the linear predictors did not capture the non-linear trend of Time. Right: GAMM model of all data, as presented in Section 4. The residuals from one specific trial are marked with white dots.

5.2 Autocorrelation of residuals

One of the most important checks, especially for time series analysis, involves inspection of the structure in the residuals. Structure in the residuals indicates that the model fails to account adequately for the structure that exists in the data. In other words, the model does not provide a very good fit of the data. A quick way of checking for structure is plotting the residuals against the fitted values, or a continuous predictor such as Time, as in Figure 9. Ideally, the residuals form a random cloud without any trends. As discussed in the previous section, the black solid line in the Right panel of the plot suggest that there is no trend left in the residuals of the GAMM model for Time values. However, the residual plot clearly shows trial structure in the residuals – sequences of residuals that seem connected. To highlight this, we have colored the residuals for one specific trial white in the right panel of Figure 9 (the same example as in Figure 4). Such structure is called *autocorrelation* in the residuals. The autocorrelation means that the value of a residual is correlated with the residual of the previous data point.

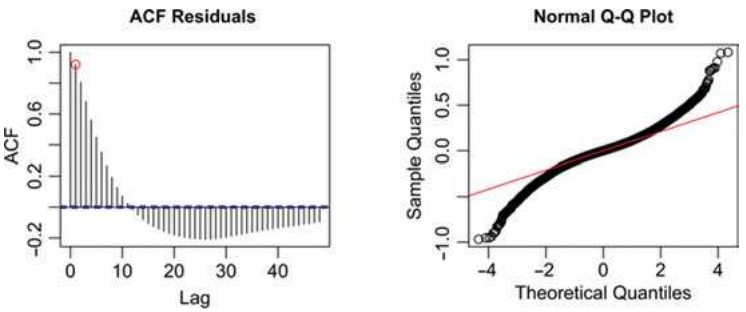

Figure 10: Residuals of the GAMM model for mouse tracking data. *Left panel:* Autocorrelation of residuals. The red circle marks the lag 1-value. *Right panel:* QQ-norm plot.

An ACF (autoregressive function) plot is used to diagnose the autocorrelation (left panel of Figure 10). On the X-axis of the plot the lag is represented, the number of trials back with which the correlation is calculated. The autocorrelation at lag 0 is necessarily 1, because this is the autocorrelation between all residuals and themselves. The autocorrelation at lag 1 (indicated with the red circle) is 0.94. So, the value of the residuals is 94% determined by the residual of the previous sample. The lag 2 value represents the autocorrelation between the residual and the residual of two samples backward. Ideally, the autocorrelation at the lags larger than 1 is as low as the blue dashed lines indicate. Autocorrelation can be described by an autoregressive model of order n, AR(n): $X_t = c + \sum_{i=1}^{n} \rho_i X_{t-i} + \varepsilon_t$, in which c is a

constant, ρ_i is the amount of autocorrelation between the residuals and the residuals at lag i, and ϵ is noise.

Autocorrelation is generally associated with time course data, in which the samples are clearly related (e.g. van Rij et al. 2019), but also can show up in behavioral data, such as reaction times (e.g. Baayen and Milin 2010). Reaction times may show learning effects (gradually getting faster as the task becomes more familiar), fatigue, and concentration fluctuations. One of the causes of autocorrelation is correlation in the sampled data. The consequence of autocorrelation is that the model reports too much confidence in the estimates, because the model works with the assumption that all data points are independent. Thus, the model reports too small confidence bands and too low p-values, and the generalizability of the model is reduced. Note that autocorrelation is not a problem specific to GAMMs, but arises with every regression method that tries to fit time series data. When fitting linear regression models on time series data the autocorrelation may be more severe as the linear regression lines cannot capture nonlinear trends over time (see Figure 9, Center panel). (The stronger autocorrelation in the residuals might be the reason why the LME polynomial model in this chapter reports significant differences for Accent and Congruency, even though these effects are not found to be significant with GAMMs.) A first step in analyzing time course data is reducing the sample size as far as possible so that the correlation between consecutive samples is reduced.

To inspect what causes the autocorrelation in the residuals of our mouse tracking data analysis, we visualize the fit of three randomly selected trials (Figure 11, Left panel). The gray lines are the raw data, the red lines the model fit (summed effects), and the gray shaded areas the residuals (difference between the data and the regression model). As time-series data by definition consist of sequences of strongly correlated measurements, the difference between the estimated regression lines and the data are strongly autocorrelated residuals. The model fits a unique line for each event, i.e., participant-trial combination, based on the *by-participant-condition* non-linear random smooth over Time and the *by-item-condition* non-linear random smooth over Time. The estimated effect is also adjusted with a random intercept for each unique event. Although the model captures the general trends of the three trials, it is not completely able to fit each individual mouse trajectory precisely. For a more precise model fit a *by-event* (unique participant-trial combination) non-linear random smooth needs to replace the current random effect structure. The Center panel of Figure 11 shows the much more precise model fit when by-event smooths are included: the residuals (gray shaded areas) are much smaller. However, autocorrelation is measured independently of the residual size: The Right panel of Figure 11 shows that the autocorrelation is reduced, but did not

disappear completely. Nevertheless, smaller residuals will reduce the consequences of the autocorrelation in the residuals. Thus, it is very important to improve the model fit.

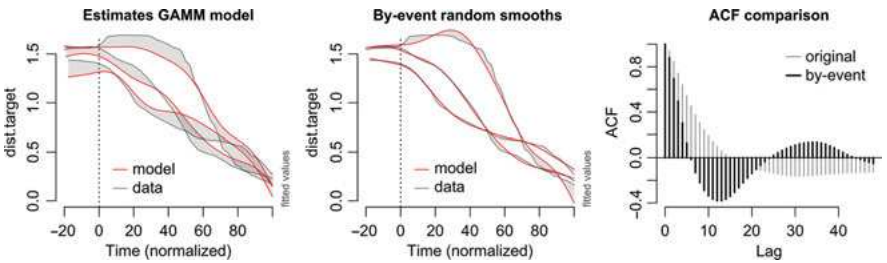

Figure 11: Data of three randomly sampled trials (black lines) compared with the model's estimates for the same trials (red lines) and the residuals (gray shaded areas). *Left panel:* fit of GAMM model discussed in Section 3c, *Center panel:* GAMM model with by-Event random non-linear smooths. The *Right panel* combines the ACF of the two models, the thin lines represent the original GAMM model, the thick lines represent the GAMM with by-Event random smooths.

Besides down sampling and improving the model fit, GAMMs as implemented in the R package *mgcv* (Wood 2011, 2017) provide another solution to account for the autocorrelation in the residuals. It is possible to include an AR(1) model (autoregressive model of order 1, as introduced above) for the residuals so that the GAMM model can take into account that the residuals are correlated while fitting the data. To include an AR(1) model, first the autocorrelation of lag 1 is estimated from a GAMM model that did not include an AR(1) model, and this value is provided to the new model as an autocorrelation measure. The model will adjust its confidence estimation accordingly. A model comparison procedure can be used to optimize the estimation of the autocorrelation parameter. Including an AR(1) model is a practical solution when the random effects structure that can be included is limited. However, the method is not perfect: an AR(1) model with the same autocorrelation parameter for all participants is often too simplistic and does not always sufficiently reduce the autocorrelation (Baayen et al. 2018; van Rij et al. 2019). Options to account for correlation in the residuals are also available in LME (package *nlme*, Pinheiro et al. 2017).

To summarize, for GAMMs analyses there are currently three solutions available to reduce the autocorrelation: (i) reducing the sample size, (ii) improving model fit by including by-event random smooths to capture individual time series,

and (iii) including an AR(1) model so that the model takes into account the autocorrelation in the model fit by reducing its confidence in the observations.

5.3 Distribution of the residuals

The distribution of the residuals is generally investigated with a QQ-norm plot which plots the distribution of the model's residuals against a theoretical normal distribution with a similar standard deviation and mean (Right panel of Figure 10). Ideally, the residuals follow a straight line, which represents the normal distribution. However, for the GAMM model of mouse tracking data we see that the residuals deviate from a normal distribution, with the lowest residuals lower than expected for a normal distribution and with the highest residuals higher than expected. This pattern suggests that the data are following a t-distribution rather than a normal distribution, because the t-distribution has heavier tails than the normal distribution, i.e. higher probability for extreme high and low values than with normal distribution. This symmetrical deviation from normality is difficult to correct with transformations.

Residuals following a t-distribution are also regularly found in other biophysiological data, such as pupillometry measures or EEG data. GAMMs (the package *mgcv* version 1.8 or higher), but not LMEs, offer the possibility to fit a scaled t-distribution to the data. A disadvantage is that running the model under the assumption of a scaled t-distribution is still relatively slow (in mgcv version 1.8–17), so it is not possible to include non-linear random smooths for predictors with many levels when using this distribution. We advise comparing the model's estimates based on a Gaussian model and based on a scaled t-distribution to see whether and how the estimates change. As the autocorrelation in the residuals seem to affect the model estimates more severely, we generally focus on reducing the autocorrelation first.

5.4 Collinearity

In an ideal world, explanatory variables would be related to our dependent variables, while being unrelated to one another. Indeed, traditional experimental design can be thought of as an attempt to bring about just this situation. This would allow theorists to maximize explained variance in the dependent variable while simultaneously working toward the most comprehensive and accurate theoretical model.

We do not live in an ideal world, though. Many potential explanatory variables are related not only to our dependent variables but to one another, and sometimes strongly so. This is true both of stimulus characteristics such as frequency, length, and concreteness, and of participants characteristics such as age, education, and reading proficiency. This situation is referred to as *collinearity* (or sometimes *multicollinearity*). *Essential collinearity* refers to the underlying structure of a dataset, while *non-essential collinearity* simply depends on the particular scales on which the variables have been measured. *Essential collinearity* is the type that researchers care about most. We will return to this distinction below, when discussing the common suggestion that mean-centering improves collinearity. For now, we simply note that mean-centering does not improve essential collinearity in any way.

Collinearity can bring with it a set of problems for researchers. One is that, if a person is using significance testing, it is possible for a statistical model to explain a significant proportion of the variance in the dependent variable without a single one of the individual predictor variables being significant. This can occur because variance that can be explained by multiple explanatory variables is not assigned to any single one of them, although it is counted as explained variance in the evaluation of the overall model.

Perhaps more unsettling for researchers is the issue of *suppression*. Many different definitions of suppression have been used, but following Wurm and Fisicaro (2014), we use the term to refer to any case in which the sign of a predictor variable's zero-order correlation with the dependent variable (i.e. the bivariate correlation, controlling for no other variables) differs from its sign in a larger analysis with multiple explanatory variables.

Friedman and Wall (2005) showed that when there are only two predictor variables, it is easy to understand and predict what will happen to the signs of the regression coefficients as a function of the strength of the correlations. On the one hand, it is comforting to know that it will always be the weaker of the two predictors that will show the sign change. On the other hand, if the zero-order correlations between each predictor and the dependent variable are similar in size to one another, then in another data set (even using the same stimuli and task) their relative sizes might reverse. This would cause the sign of the previously-larger effect to be the one that now changes. In addition, such effects become harder to understand and predict with each explanatory variable added to the model.

The troubling effects of even slight changes in these "initial conditions" are what make some researchers mistakenly assert that there is computational instability in the models. Friedman and Wall (2005) say "multicollinearity does not affect standard errors of regression coefficients in ways previously taught"

(p. 127), and provide a very nice demonstration that any "instability" is not computational. It has to do with the underlying correlational structure of the dataset.

In the next section we will highlight some of the strengths and weaknesses of some of the potential solutions that have been offered.

Residualizing. Residualizing is a technique in which one predictor variable is regressed on one or more other predictor variables. The residuals (i.e. the unexplained portion of the variance) are retained and used in place of the original predictor variable. By definition this residualized variable will be uncorrelated with any variable on which it was residualized, so this method appears to offer a useful solution to collinearity. However, Wurm and Fisicaro (2014) present evidence from the literature that the risk of misinterpretation far outweighs anything that might be learned from such analyses, particularly because any information available from such analyses is also available from methods far less likely to be mischaracterized or misunderstood. In addition, for complex situations like those found in actual psycholinguistic studies, the likelihood increases that an analysis including residualized predictors cannot be meaningfully interpreted at all.

There is also a general interpretational problem that comes with residualization. This is illustrated nicely by Breaugh's (2006) example based on the strong correlation between the heights and weights of professional basketball players. He found that players' heights predicted their rebounding totals only if their weights were not controlled for. He questions, though, how one might interpret a height variable from which weight has been residualized. He says that " . . . making subjunctive statements based upon a residual variable is inappropriate. Simply stated, there is no basis to assume that, if in reality height and weight were uncorrelated, height would not be related to rebounds. Given they are correlated, and highly so, we simply have no way of knowing" (p. 439). In the long run we are better off trying to understand why the predictors are correlated, which of course is easy for the present example.

Principal components. An alternative approach is to perform a principal components analysis on the set of predictor variables one wishes to use. Several methods exist, but in general the idea is that the number of predictors will be reduced to a small number of principal components that are orthogonal (i.e. uncorrelated with one another). The drawback is that the original predictor variables are now gone, and all we have left are mixtures of the predictors that cannot be analyzed back into their constituent parts. One can sometimes make statements such as "Principal Component #1 seems to be related to word frequency" by examining how individual predictors correlate with it, but in general that will not be sufficient for development or testing of a theoretical model.

Baayen, Wurm and Aycock (2007) used principal components analysis to capture sequential dependencies in a trial-by-trial analysis of lexical decision times. It was probably harmless to use in this situation, because they did not care about recovering the structure of the original predictors, but as Wurm and Fisicaro (2014) showed, if the only concern was in removing that extraneous variance (or "controlling for" it), then the approach bought them nothing.

Mean-centering. A number of researchers have suggested mean-centering (i.e. subtracting from each score the mean on that variable) as a way to reduce collinearity. Mean-centering addresses non-essential collinearity for the simple reason that it changes the scaling of the variables, but unfortunately it does nothing whatsoever to address the underlying structural relationships between the variables. Thus, essential collinearity is left unchanged by mean-centering. Worse still, mean-centering can mask some of the diagnostics used to assess collinearity (Belsley 1984; Pedhazur 1997), leading researchers to the mistaken belief that they have solved the problem. A number of authors, including Dalal and Zickar (2012), nevertheless recommend mean-centering because it can make the interpretation of regression coefficients easier and more immediately meaningful, but it does not in any way improve essential collinearity.

Other approaches include some that compute solutions over many different permutations and/or combinations of predictor variables. One example of this is random forests (Breiman 2001; Strobl, Malley and Tutz 2009), which assign a higher importance to a predictor variable if its original version predicts the dependent variable much better than a permuted version does. One practical concern is that even with current computing power, the analyses can take several hours to run (Tagliamonte and Baayen 2012). An additional question that has not been the topic of research so far as we are aware is the sensitivity of random forest computations to the "initial conditions" we spoke of above. That is, if predictor X_1 has a slightly stronger relationship to the dependent variable than predictor X_2 does, will it necessarily emerge as the more important predictor across the summary of the permuted analyses? If so, researchers are in the same worrisome situation of having to decide whether that initial ordering of the variables reflects reality, or whether it is perhaps something idiosyncratic about the particular dataset being analyzed.

Ridge regression. A final approach we will mention is called ridge regression (Hoerl 1962). It prevents error variance from increasing under conditions of high collinearity, and produces slightly conservative parameter estimates. The biggest drawback in our view is that it cannot be used with the kinds of designs most frequently employed by psycholinguists (repeated-measures designs, which are usually being analyzed with multilevel or mixed-effects models). It can, however, be used to analyze item sets.

We believe it worth emphasizing that all of these approaches will fail in one respect or another. Darlington (1990) wrote that it is a "misconception about collinearity ... that more advanced statistical methods might someday eliminate the problem. The problem is essentially that when two variables are highly correlated, it is harder to disentangle their effects than when the variables are independent. This is simply an unalterable fact of life" (Darlington 1990: 131; see also Darlington 1968; Pedhazur 1997).

Suggestions. Most textbooks on regression (e.g. Tabachnick and Fidell 2007) contain recommendations for what one might do to deal with high collinearity. Such suggestions include things like creating composite variables, omitting some predictors, and doing nothing. Indeed, if our goal for a particular set of predictor variables is simply to explain variance, then the best approach is to include any and all predictors that might have a relationship with the dependent variable. The same holds true if our goal is simply to be in a position to say that we have "controlled for" the effects of one or more variables. We can safely put them into our models and go about our business without any concern for what might have happened to their signs, or their p-values. In many cases, though, this won't do. No researcher is willing to maximize explained variance at the expense of parsimony and coherence in their theoretical model. If the goal is to have a good theoretical model, then we're back to having to decide what to do.

We would like to offer the suggestion that whatever approach is taken skirts the real issue. Statistical "control" (and everything that means: residualizing, principal components analysis, random forests, even the whole idea of multiple regression itself) is an attempt to "equate" or balance stimuli, which we talked about above in the context of traditional factorial designs. Meehl (1970: 385) spoke eloquently about the difficulties this poses: "When a social scientist of methodological bent tries to get clear about the meaning, proof, and truth of those counterfactuals that interpret statistical formalisms purporting to 'control the influence' of nuisance variables, he is disappointed to discover that the logicians are still in disagreement about just how to analyze counterfactuals" (see also Campbell, Converse and Rodgers 1976). Anderson (1963: 170) was more to the point a few years earlier: " ... one may well wonder exactly what it means to ask what the data would be like if they weren't what they are."

Darlington (1990: 155) says that "suppression rarely occurs in real data". Cohen et al. (2003) assert that it is more common in fields like economics than in the social sciences, because in those fields variables can sometimes have "equilibrium-promoting effects." We think it likely, though, that Darlington and Cohen et al. did not foresee the kind of statistical models being run in

modern psycholinguistics, which can sometimes contain literally dozens of interrelated predictors.

Such models are probably indefensible anyway, and thus force us to confront the possibility that regression-based techniques are not up to the task we are asking of them. At some point a researcher must confront more directly what all of these intercorrelations mean, instead of hoping for a new, more creative analytic strategy to emerge. Why do these things co-vary? Which one might have temporal or theoretical priority? Which model is the most useful, not only in terms of explaining this dataset but in terms of making predictions about other datasets? Whatever approach or combination of approaches is used, we would urge researchers to use clear, precise, and proper language, and to include as much information as possible for those interested in replicating the analyses.

6 Discussion

In this chapter we have outlined three different analysis methods that could be used for analyzing psycholinguistic data. All three methods aim to account for the variability between participants and stimuli, which characterizes psycholinguistic data. However, the methods each have their own strengths and weaknesses.

6.1 Choosing a statistical method

Repeated-measures ANOVA is the oldest method and still most commonly used. The method provides robust results for balanced factorial designs without missing data, and with a dependent variable that is normally distributed and with the variance being homogeneous across conditions. Advantages of this method are that it is well-documented and that the results are easy to report. Disadvantages are that several analyses are required, i.e. F1 test, F2 test, and post-hoc tests for interpreting the results, and that the method is fairly limited in use. In practice, psycholinguistic data often contains covariates, such as frequency, time, or age, and missing data is a common problem with human participants or corpus data. The method is not suited for analyzing time series data, because it does not allow inspection of the time course directly, but rather requires collapsing over time windows.

Linear mixed-effects modeling (LME) is a well-established alternative analysis for repeated-measures ANOVA. The method is more robust than ANOVA with

missing data and can handle unbalanced data and those that are not normally distributed, such as binomial data or count data. An advantage of the method is that the method allows direct analysis of sample data without the need to average. This reduces the number of analyses to perform. Other advantages are the possibility to include covariates, the flexibility of the method with missing data and unbalanced designs, and the interpretation of the results. The interpretation of the results is relatively easy in comparison with the other discussed methods, because the method provides the estimated coefficients. No additional post-hoc tests are required and the estimated variability between participants and items can be easily inspected. Disadvantages of the method are that it requires more time to run the analysis, and that the method is more vulnerable to anti-conservative estimates, i.e. over-fitting the data, when the random effect structure is too limited (Barr et al. 2013). It requires more effort to determine the structure of the random effects, because the maximal random effect structure (i.e. subjects and items vary in their sensitivity to all experimental manipulations), is not always possible (Baayen et al. 2017). Another disadvantage is that it cannot handle non-linear covariates very easily, as the shape of the non-linear pattern needs to be specified by the user. Finally, there are not many possibilities to account for the autocorrelation problem.

Generalized additive mixed modeling (GAMM) is a relatively new *non-linear* mixed-effects regression method that is particularly suited for analyzing non-linear data, such as time series data or data with non-linear covariates. It shares with LME that the method can handle unbalanced data and not normally distributed data, such as binomial data or count data, and allows direct analysis of sample data without the need to average. One of the main advantages of GAMMs is a better understanding of the data, because the method relies much more than the other methods on visualization of the estimates and results. Other advantages are the possibility to include non-linear effects and interaction surfaces, and non-linear random effects, and the possibility to account for autocorrelation in the residuals with an AR1 model. Disadvantages of GAMMs are that they can require a long time to run, and that the interpretation of the model takes more time, because the non-linear effects need to be visualized as coefficients are not provided. Another disadvantage of GAMMs is that finding the best-fitting model is less straightforward than with LME, as models are not strictly nested. A final disadvantage is that the results are less generalizable when autocorrelation plays a role, or when the model does not fit the data very well, for example when only limited random effects can be included. In these cases, one needs to be cautious with the interpretation of the results.

Thus, these three methods could be considered complementary: to analyze the data from a simple factorial balanced design, it may be valid to use a

repeated-measures ANOVA although mixed-effects models are equally powerful alternatives (Baayen 2008: Chapter 7); but for unbalanced designs or data sets with linear covariates LME is a better choice, and when non-linear effects play a role GAMM is the preferred option. So instead of focusing on one single analysis method and letting that particular analysis method determine the design of our experiments, as often seem to be the underlying reason for factorial designs, the mixed-effects methods provide us a powerful tool to investigate different questions using more flexible designs. For example, when we only have ANOVA available as statistical method we need to carefully control the frequency of our stimuli in equally high and low frequency words for our different manipulations, dichotomizing frequency. However, GAMMs allow us to sample words with a range of frequencies and include frequency as continuous measure in our analysis. When we would like to use a GAMM analysis it is actually better to sample words with different frequencies from a *range* instead of selectively choosing the words with low and high frequency. In other words, the statistical methods that we have available for use will influence the choice of design. Moreover, the statistical method will also shape the questions we ask: for example, non-linear regression methods allow us to ask *at which moment* in the time course two conditions start to differ, instead of *whether* we detect early and/or late differences.

6.2 Implications for design

ANOVA compatible designs in reaction time studies have long dominated the analytic landscape in psycholinguistics as well as in other domains of inquiry (for a review, see Van Zandt 2002). Many did and still do believe that the only competent methodology is an experiment with a factorial design that allows for hypothesis testing and causal inference. By implication a study with a correlational design is necessarily inferior because it describes only an association. Assumptions like these motivate a common research practice in the domain of psycholinguistics that is to treat continuous measures dichotomously, by sampling at two points (ranges) along a continuum, and then matching the means of those groups on other relevant factors. Data generated in this framework are subject to several shortcomings, which include the consequences of (i) control by matching (ii) control by counterbalancing (iii) limitations of analyzing means and (iv) diminishing the richness of big data.

i) Control by matching. In the traditional design, it is typical to "manipulate" an independent variable or two of interest and then "match" words across the various levels of other potentially relevant measures. For example, it is

typical to manipulate target frequency in a factorial manner (e.g. a range for high and a non-overlapping range for low treatment conditions) and control word length and number of words that differ from the target by one letter or phoneme (neighbors). Here, "control" entails dichotomizing a continuous variable and then matching means for each group along those other possible measures. One obvious problem with matching a measure of central tendency is that it does not make the distributions that they describe comparable. Parametric statistical analyses work best when distributions do not have long tails and outliers. Matching only on a measure of central tendency can violate this assumption. In part, the consequences of imposing a dichotomous structure on a continuous measure depend on the non-linearities in its behavior (see Baayen 2010).

When two independent variables are manipulated factorially, matching means across combinations of levels or treatment conditions gets even more tedious. The problem gets more complex when the measures to be matched are correlated, for example word length and frequency. Shorter words tend to be higher in frequency (*the*, *and*, *his*, *her*) and, because these covary, the set of words that are short but low in frequency (e.g., *awl*, *cob*, *ewe*) will, by definition tend to be statistically atypical. More realistically, clusters rather than pairs of measures tend to be related. For example, the many measures of frequency tend to be related not only to measures of length but also to measures of form similarity captured by neighbors. Therefore, manipulating frequency while matching on number of neighbors and length requires breaking a natural co-variation. One obvious implication is that words are not randomly selected to fill out a factorial design that includes measures that covary. The practical consequence is that matching in this way is likely to lead to selecting low frequency words that are atypically non-homogenous on related measures like number of neighbors or perhaps bigram structure. For example, whereas short words tend to have many neighbors, orthographic neighbors for *awl*, *cob*, *ewe* are 4, 28 and 5, respectively. The severity of the matching problem depends on the degree of correlation among measures. It is a general problem and applies not only to correlations of word frequency with form described above or of word frequency with semantic measures such as wordliness or semantic density (Keuleers and Marelli, this volume). The theoretical implication of reliable interactions such as these is that the conventional interpretation of frequency, tying it to activation of lexical entries without regard to their constituents may be flawed (Kuperman et al. 2009). This cannot be evaluated with factorial designs, however.

In this example, frequency, which is by its nature a continuous predictor, is treated dichotomously. Analyses of covariance provide a modest remedy when the focus is only a select number of measures and the correlation among

them is not strong. Nonetheless, these analyses assume a linear relation between predictors. At least for frequency, linearity cannot be assumed without careful inspection of the dataset.

ii) Control by counterbalancing. In factorial designs, control is typically based on random assignment of participants and sometimes items to conditions along with changing the order in which items are presented or the location at which they appear on the screen. The underlying assumption is that counterbalancing assignment and order or location is sufficient to alleviate random differences between participants and between items. While in practice order effects such as training or fatigue are not always removed by aggregation, it is generally assumed that any effect worth studying should be robust to the noise associated with trial number or sequential order. Counterbalancing in this manner makes it basically impossible to track behavior that changes during the course of the experimental session as well as interactions that involve differences between participants or items. For example, with relevant controls, skilled readers tend to perform more consistently during the course of an experimental session than do less skilled readers. This cannot be detected easily when skill is treated dichotomously. Similarly, evidence that participants catch on or otherwise adjust to a property that differs among words (e.g. native or non-native accent) as they progress through the experimental session or trial would be missed. Finally, analyses include only correct trials therefore performance on prior trials is likewise treated as noise.

iii) The implications of aggregating over participants or items. Along with counterbalancing in this way is the convention of using means by participant by condition or of word by condition as the unit of analysis. For a period, it was conventional to report sets of analyses, one with subjects as the random effect and a second with items (Clark 1973). The rationale was to demonstrate that the findings generalize beyond the sample of participants and language materials that were tested. The fact that participants were nested within a particular combination of items and conditions was ignored (e.g. Raaijmakers et al. 1999). The fact that some participants perform more poorly than others and contribute fewer correct data points to their mean for a condition was also ignored when means are the unit of analysis. This practice becomes particularly problematic when missing data are meaningful as with clinical populations or studies that track acquisition (Keuleers and Marelli 2020, this volume).

iv) Diminishing the richness of big data. Large-scale datasets compiled from human behavioral measures (eye tracking, EEG), linguistic corpora (Nelson association norms, CELEX) or collected from digital social media provide data about individuals and about groups. New technologies have made salient many of the inadequacies of the factorial approach, especially with respect to changes

in behavior over time and have inspired the adaptation of new quantitative analyses and measures in non-linguistic as well as linguistic domains. One now classical way to reduce the dimensionality of these data is by focusing on peaks and where they arise relative to the onset of an event. At its simplest, this technique assumes that one can identify a peak and distinguish it from a prolonged elevation and, that it is possible to define a peak globally rather than relative to a local baseline. With these constraints, it becomes more complex to detect a peak in conjunction with a general drift toward lower values or other types of artifacts. Of crucial importance with many of the technologies is appreciation of how behavior changes over time. There are multiple techniques of varying complexity to incorporate variability over time.

The simplest is to define bins or other fixed intervals and revert to computing means over smaller intervals. Decisions as to how many bins are often made on an ad hoc basis with little consideration of what makes one smoothing procedure preferable to another (e.g. detecting possible non-linear patterns). At the same time, choice of procedure can have dramatic consequences for the outcomes that emerge and the interpretation they warrant. For example, analyses of reaction time studies based on movement of a mouse to one of two designated locations on the computer screen depending on the decision on individual trials (audio and visual match, audio and visual mismatch) could restrict the dependent measure to time to execute the mouse trajectory from beginning to end. Alternatively, the analysis could divide the average trajectory into a number of smaller trajectories and then focus either on those means (x and y coordinates) or on how those means change over steps. Of course, one could also look at time to initiate the movement. Conditions could yield comparable total reaction times with different onsets to movement in which case we would know that, on average, participants who started later moved the mouse faster. Similarly, we could ask whether those who moved the mouse faster tended to have a more curved trajectory than those who moved it more slowly. Obviously, incorporating time steps into an analysis increases the number of dependent measures one can examine but restricting the analysis to means per time step dramatically diminishes the richness of the data.

6.3 Assessing significance

On the other hand, the analysis of more naturalistic but less balanced data will also reveal the limitations of the statistical techniques available. Problems such as limited sample sizes, non-normally distributed residuals, autocorrelation, and collinearity result in less reliable p-values, and less coherent model

comparison procedures. Therefore, we strongly advise using different methods to test whether the experimental manipulations really explain variance in the data. Different methods that apply to all mixed-effects models are (i) a careful model comparison procedure to select the best-fitting model (this can be done manually, but there are also packages available that implement automatic comparison procedures), (ii) inspection of the model summaries and random effects, and (iii) visualization of the model's estimation of effects. Visualization of the model's estimates is traditionally not used so much in statistical analysis. However, the more complex the model the more important visualization is. The visualization of model estimates will quickly reveal problems with the model fit, for example by not capturing subject variability or by outliers that drive the significance of effects, and will aid the interpretation of the results.

If these three sources of information do not converge to the same conclusion, it is useful to investigate why this might be the case. The lack of convergence basically signals that the model's results are not stable, which could be due to one of the earlier described problems. In addition to these model selection methods, we strongly encourage investing time in model evaluation. Inspection of the residuals and testing the assumptions of regression models reveals critical information with respect to the generalizability and interpretation of the results.

In this chapter we have emphasized that the purpose of statistical analysis is not generating p-values, but to model the data to distinguish accidental patterns from replicable effects. We argue that when we want to take advantage of the recent experimental techniques to investigate online language processing such as eye tracking, EEG, articulography, or mouse tracking, we need to *understand* the patterns in the data instead of *reducing* and *simplifying* these patterns in order to derive a p-value. In this perspective, the limitations of the statistical model provide useful information that help us to understand the data.

6.4 Summary of results

In a repeated-measures ANOVA based on participant and item means for each condition, participants performed at chance level in recognizing the accent at study for words tested in an English accent. For the words in a Chinese accent, however, the participants' responses show a clear effect of study-test congruency. In all analyses target location (right versus left) was counterbalanced and differences due to location were treated as noise because they were not linguistically meaningful. LME analyses permitted the introduction of random slopes and intercepts and revealed that participants differed in overall performance (RT, accuracy) and in whether they treated match and mismatched study-test congruency trials in

the same manner. GAMMs, a non-linear regression analysis allowed us to ask whether participants differed as they moved the mouse right or left during an experimental trial. It can account for hesitations and abrupt shifts in the mouse trajectory with the target distance measure and results can be interpreted as indices of uncertainty and changed decisions (Freeman, Dale and Farmer 2011; Freeman and Johnson 2016). Longer duration responses by mouse movements followed initially a more direct path than the short duration responses, but deviated in the mid portion of the trajectory – indicating uncertainty or revision of the response. More interestingly, some but not all participants used the additional time to produce a relatively straighter path. GAMMs are preferable to LME with polynomial curves because they specify the requisite polynomial and permit the inclusion of non-linear interaction surfaces. It is possible to determine at which moment the trajectories of different conditions start to differ. With respect to effects of test-study congruency, the GAMM model found no differences between an American and a Chinese accents.

References

Anderson, Norman H. 1963. Comparison of different populations: Resistance to extinction and transfer. *Psychological Review* 70(2), 162–179.

Baayen, R. Harald. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge (UK): Cambridge University Press.

Baayen, R. Harald. 2010. A real experiment is a factorial experiment. *The Mental Lexicon* 5 (1) 149–157.

Baayen, R. Harald, Douglas J. Davidson, & Douglas M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59 (4). 390–412.

Baayen, R. Harald, & Petar Milin. 2010. Analyzing reaction times. *International Journal of Psychological Research* 3 (2). 12–28.

Baayen, R. Harald, Jacolien van Rij, Cecile de Cat, & Simon N. Wood. 2018. Autocorrelated errors in experimental data in the language sciences: Some solutions offered by Generalized Additive Mixed Models. In Speelman, D., Heylen, K. and Geeraerts, D. (eds), *Mixed Effects Regression Models in Linguistics*, 49–69. Berlin, Springer.

Baayen, R. Harald, Shravan Vasishth, Reinhold Kliegl, & Douglas M. Bates. 2017. The cave of Shadows. Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language* 94. 206–234.

Baayen, R. Harald, Lee H. Wurm, & Joanna Aycock. 2007. Lexical dynamics for low-frequency complex words: A regression study across tasks and modalities. *The Mental Lexicon* 2 (3). 419–463.

Barr, Dale J., Roger Levy, Christof Scheepers, & Harry Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language* 68 (3). 255–278.

Bates, Douglas, Martin Mächler, Ben Bolker, & Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67 (1). 1–48.

Bates, Douglas, Reinhold Kliegl, Shravan Vasishth & R. Harald Baayen. 2015. Parsimonious mixed models. arXiv:1506.04967, 1–21.

Belsley, David A. 1984. Demeaning conditioning diagnostics through centering. *The American Statistician* 38. 73–77.

Breaugh, James A. 2006. Rethinking the control of nuisance variables in theory testing. *Journal of Business and Psychology* 20 (3). 429–443.

Breiman, Leo. 2001. Random forests. *Machine learning* 45 (1). 5–32.

Browne, Michael W. 2000. Cross-validation methods. *Journal of Mathematical Psychology* 44. 108–132.

Campbell, Angus, Philip E. Converse, & Willard L. Rodgers. 1976. *The quality of American life: Perceptions, evaluations, and satisfactions*. New York: Russell Sage Foundation.

Clark, Herbert H. 1973. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior* 12 (1973). 335–359.

Cohen, Jacob, Patricia Cohen, Stephen G. West, & Leona S. Aiken. 2003. *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Dalal, Dev K., & Michael J. Zickar. 2012. Some common myths about centering predictor variables in moderated multiple regression and polynomial regression. *Organizational Research Methods* 15 (3). 339–362.

Darlington, Richard B. 1968. Multiple regression in psychological research and practice. *Psychological Bulletin* 69. 161–182.

Darlington, Richard B. 1990. *Regression and linear models*. New York: McGraw-Hill Publishing Company.

Fang, Yixin. 2011. Asymptotic Equivalence between Cross-Validations and Akaike Information Criteria in Mixed-Effects Models. *Journal of Data Science* 9. 15–21.

Freeman, Jonathan B., & Nalini Ambady. 2010. MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behavior Research Methods* 42. 226–241.

Freeman, Jonathan B., Rick Dale, & Thomas T. Farmer. 2011. Hand in motion reveals mind in motion. *Frontiers in Psychology* 2. https://www.frontiersin.org/articles/10.3389/fpsyg.2011.00059/full

Freeman, Jonathan B., & Kerri L. Johnson. 2016. More than meets the eye: split-second social perception. *Trends in cognitive sciences* 20 (5). 362–374.

Friedman, Lynn, & Melanie Wall. 2005. Graphical views of suppression and multicollinearity in multiple linear regression. *The American Statistician* 59 (2). 127–136.

Gelman, Andrew, & Jennifer Hill. 2007. *Data analysis using regression and multilevel hierarchical models*. Cambridge: Cambridge University Press.

Hoerl, Arthur E. 1962. Application of ridge analysis to regression problems. *Chemical Engineering Progress* 58. 54–59.

Hothorn, Torsten, Frank Bretz, & Peter Westfall. 2008. Simultaneous Inference in General Parametric Models. *Biometrical Journal* 50 (3). 346–363.

Keuleers, Emmanuel and Marco Marelli. 2020. Resources for mental lexicon research: A delicate ecosystem. In Vito Pirrelli, Ingo Plag & Wolfgang Dressler (eds.), Word

Knowledge and Word Usage: a Cross-disciplinary Guide to the Mental Lexicon, 164–184. De Gruyter.
Kuperman, Victor, Rob Schreuder, Raymond Bertram, & R. Harald Baayen. 2009. Reading of polymorphemic Dutch compounds: Towards a multiple route model of lexical processing. *Journal of Experimental Psychology: HPP* 35. 876–895.
Lawrence, Michael A. 2016. *ez: Easy Analysis and Visualization of Factorial Experiments*. R package version 4.4-0.https://CRAN.R-project.org/package=ez
Lin, Xihong, & Daowen Zhang. 1999. Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society (B)* 61 (2). 381–400.
Matuschek, Hannes, Reinhold Kliegl, Shravan Vasishth, R. Harald Baayen and Douglas Bates. 2017. Balancing Type I Error and Power in Linear Mixed Models. *Journal of Memory and Language* 94. 305–315.
Meehl, Paul E. 1970. Nuisance variables and the ex post facto design. In Michael Radner & Stephen Winokur (eds.), *Analyses of theories and methods of physics and psychology*, 373–402. Minneapolis: University of Minnesota Press.
Mirman, Daniel. 2014. *Growth Curve Analysis and Visualization Using R*. Boca Raton: Chapman and Hall.
Mirman, Daniel, James A. Dixon, & James S. Magnuson. 2008. Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language* 59 (4). 475–494.
Pedhazur, Elazar J. 1997. *Multiple regression in behavioral research*. Fort Worth, TX: Harcourt Brace & Co.
Pinheiro, José, & Douglas Bates. 2000. *Mixed-effects models in S and S-Plus*. Springer, New York.
Pinheiro, José, Douglas Bates, Saikat DebRoy, Deepayan Sarkar, & R Core Team. 2017. *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-131. https://CRAN.R-project.org/package=nlme
Raaijmakers, Jeroen G., Joseph M. C. Schrijnemakers, & Frans Gremmen. 1999. How to deal with "the language-as-fixed-effect fallacy": Common misconceptions and alternative solutions. *Journal of Memory and Language* 41 (3). 416–426.
Roberts, Seth, & Harold Pashler. 2000. How persuasive is a good fit? A comment on theory testing. *Psychological Review* 107. 358–367.
Strobl, Caroline, James Malley, & Gerhard Tutz. 2009. An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods* 14(4). 323–348.
Tabachnick, Barbara G., & Linda S. Fidell. 2007. *Using multivariate statistics* (5th ed.). Boston: Pearson Education, Inc.
Tagliamonte, Sali, & R. Harald Baayen. 2012. Models, forests and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change* 24. 135–178.
Vaci, Nemanja, Bartosz Gula, & Merim Bilalić. 2015. Is age really cruel to experts? Compensatory effects of activity. *Psychology and Aging* 30. 740–754.
van Rij, Jacolien, Petra Hendriks, Hedderik van Rijn, R. Harald Baayen, & Simon N. Wood. 2019. Analyzing the time course of pupillometric data. *Trends in Hearing Science* 23. 1–22.
van Rij, Jacolien, Martijn Wieling, R. Harald Baayen, & Hedderik van Rijn. 2017. *itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs*. R package version 2.3. https://CRAN.R-project.org/package=itsadug

Van Zandt, Trisha. 2002. Analysis of response time distributions. In John T. Wixted (Vol. Ed.) & Hal Pashler (Series Ed.) *Stevens' Handbook of Experimental Psychology (3rd Edition)*, *Volume 4: Methodology in Experimental Psychology*, 461–516. New York: Wiley Press.

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

Wood, Simon N. 2017. *Generalized additive models: An introduction with R*. Boca Raton: Chapman and Hall. Second Edition.

Wurm, Lee H., & Sebastiano A. Fisicaro. 2014. What residualizing predictors in regression analyses does (and what it does *not* do). *Journal of Memory and Language* 72. 37–48.