Bachelor Thesis

Mixed Model Analysis of Trial History in Naming Experiments

Maximilian Michael Rabe

Universität Potsdam

Student number: 766774

maximilian.rabe@uni-potsdam.de

rabe.maximilian@gmail.com

Date: September 4, 2015

Supervisor

Prof. Reinhold Kliegl, Department Psychologie, Universität Potsdam

Co-Supervisor

Prof. Michael E. J. Masson, Department of Psychology, University of Victoria

Abstract

Mixed Model Analysis of Trial History in Naming Experiments

by

Maximilian Michael Rabe

Submitted to the Examination Board in partial fulfilment of the requirements for the Bachelor of Science in Psychology

University of Potsdam

September 4, 2015

Several authors highlighted that the time course of an experiment itself could have a substantial influence on the interpretability of experimental effects. Since mixed effects modeling had enabled researchers to investigate more complex problems with more precision than before, two naming experiments were conducted with college students, with and without non-words intermixed, and analyzed with regard to frequency, quality, interactive and trial-history effects. The present analyses build on and extend the Bates, Kliegl, Vasishth, and Baayen (2015) approach in order to converge on a parsimonious model that accounts for autocorrelated errors caused by trial history. For three of four cases, a history-sensitive model improved the model fit over a history-naïve model and explained more deviance. In one of these cases, the herein presented approach helped reveal an interaction between stimulus frequency and quality that was not significant without a trial history account. Main and joint effects, limitations, as well as directions for further research, are briefly discussed.

Keywords: Autocorrelations, Mixed Effects Modeling, Trial History, Reading Aloud

Zusammenfassung

Mixed Model Analysis of Trial History in Naming Experiments

von

Maximilian Michael Rabe

Eingereicht beim Prüfungsamt als Abschlussarbeit für den Bachelor of Science in Psychologie

Universität Potsdam

4. September 2015

Verschiedene Autoren haben darauf aufmerksam gemacht, dass bereits der zeitliche Verlauf eines Experiments einen wesentlichen Einfluss auf die Interpretierbarkeit experimenteller Effekte haben kann. Nachdem gemischte Modelle der Wissenschaft ermöglichten, komplexere Fragestellungen mit höherer Präzision als zuvor zu untersuchen, wurden zwei Naming-Experimente mit Collegestudenten durchgeführt, je mit und ohne Pseudowörter, sowie hinsichtlich ihrer Auftrittshäufigkeits-, Stimulusqualitäts-, Interaktions- und Experimentalverlaufseffekte untersucht. Die vorliegenden Analysen beruhen auf dem Ansatz von Bates, Kliegl, Vasishth und Baayen (2015) und erweitern diesen, um ein Parsimonious Model zu bestimmen, welches durch den Experimentalverlauf hervorgerufene Autokorrelationen berücksichtigt. In drei von vier Fällen verbesserte die verlaufsabhängige Analyse die Modellanpassung gegenüber der gewöhnlichen verlaufsunabhängigen Variante und klärte somit mehr Abweichung auf. In einem dieser Fälle half der Ansatz, eine Interaktion zwischen Auftrittshäufigkeit und Stimulusqualität aufzudecken, die ohne Berücksichtigung des Experimentalverlaufs nicht signifikant gewesen war. Haupt- und Interaktionseffekte, Einschränkungen sowie Anregungen für weiterführende Forschung werden kurz erörtert.

Schlüsselwörter: Autocorrelations, Mixed Effects Modeling, Trial History, Reading Aloud

Acknowledgment

I would like to express my gratitude to my supervisors Reinhold Kliegl and Michael Masson for all their guidance and help throughout my internship and this resulting thesis and who introduced me into the world of psychological research. Over the past few months I have learned more than in any class I attended. For an introduction into generalized additive models and helpful advice for the statistical analyses I cordially thank Hannes Matuschek (Universität Potsdam).

Moreover I would like to thank Sara Stengel, Alexander Thoß, and Catherina Lugauer for their helpful and inspiring remarks on this work, everyone who participated in these experiments, and all the helping hands involved in data collection, especially Marnie Jedynak (University of Victoria), thereby laying the foundation for this thesis. Last but not least, I thank my loved ones who encouraged and supported me to take this first step towards a hopefully instructive and productive journey.

Table of Contents

List of Tables

List of Figures

General Introduction

One of the most fascinating and popular topics in cognitive psychology is the exploration and modeling of cognitive processes during reading. Sternberg (1969) introduced the general assumption that two factors shown to have no significant interaction likely reflect an underlying process that occurs in distinct stages. Subsequently, the stimulus quality and frequency (familiarity) of a word were found to be additive in lexical decision (Stanners, Jastrzembski, & Westbrook, 1975; Yap & Balota, 2007) and naming tasks (Besner, O'Malley, & Robidoux, 2010; Bonin, Roux, Barry, & Canell, 2012; Carello, Lukatela, Peter, & Turvey, 1995; O'Malley & Besner, 2008). Based on the Sternberg (1969) paradigm and missing evidence for an interaction –and thereby assumed additivity– of frequency and quality, many researchers conclude that lexical processing in many cases occurs in stages, and that frequency and quality influence each one of those stages (e.g., Borowsky & Besner, 1993, 2006). In both lexical decision and naming tasks, low-frequency (less familiar) words as well as degraded (low stimulus quality) words typically lead to slower reaction times (Besner et al., 2010; Carello et al., 1995; Masson & Kliegl, 2013; O'Malley & Besner, 2008; Stanners et al., 1975).

Models of Interactive and Additive Effects

In the past decade, the joint effect of frequency and quality in naming experiments was found to be modulated by the presence or absence of non-words in the item list (Besner et al., 2010; O'Malley & Besner, 2008): In the condition where only words had been presented, low-frequency words were less affected by stimulus quality; in the condition with non-words intermixed, that joint effect was additive.

A quite successful computational model of lexical processing that was able to explain many of those effects, is the new connectionist dual process model (CDP+ model) by Perry, Ziegler, and Zorzi (2007) which predicts response times and error rates for the pronunciation of word and non-word stimuli. According to this model, visual word recognition is conducted in two independent processes. Whereas the *lexical route* has an unthresholded stimulus quality normalization module and produces responses for word stimuli (known words), the *nonlexical route* is

thresholded and produces responses for non-word stimuli (unknown words) more effectively. The faster process generates the final response.

If non-words are present, this is thought to impair or deemphasize the lexical route which is inefficient for non-words. Since this basically leaves the thresholded nonlexical route, differences in quality can be compromised at the letter level before being passed on. Under certain circumstances this creates additivity of quality and frequency effects. This is typically seen in naming experiments with non-words intermixed (Besner et al., 2010; Bonin et al., 2012; Carello et al., 1995; O'Malley & Besner, 2008).

Mixed Effects Modeling

It should be noted that it is practically impossible to actually *prove* additivity because this would premise to find evidence for the absence of a particular effect. On the contrary, there can only be significant evidence *for* an effect but not *against* it. In real-life data, there is usually always a joint effect between two effects to some degree. The actual question is: How likely is the observed magnitude of said effect different from zero or is it more likely just due to error? With the rise of improved statistical tools research has been enabled to study problems of higher complexity with more precision and account for more sources of variance at once instead of performing separate independent post-hoc analyses.

Making statistical sense of observations with the help of analysis of variance (ANOVA) had been the status quo for many years until theories and evidence became so comprehensive that more detailed analytical tools were needed. A major limitation of ANOVA is the restriction to one random factor, typically Subject. Certainly subjects do differ in how they process information and react to stimuli, but in certain fields such as psycholinguistics, also items have certain individual characteristics that likely come into play when presented in an experimental setting.

Of course we expect that the material used for an experiment has an impact on the outcome variable, e.g. number of distinctive features on detection time in visual search tasks (Treisman & Gelade, 1980). However, when the stimuli presented do not differ in clearly contoured features but are rather naturally composed like subjects, then Item indeed constitutes a second random factor. That type of variance is not to be ignored and thus research has adapted this idea by

introducing mixed-effects modeling which allows crossed random factors in one model (Bates, Maechler, Bolker, & Walker, in press; Kliegl, Wei, Dambacher, Yan, & Zhou, 2010). Random factors are mostly nominal and their levels are drawn directly from the population, which is why they are not experimental factors in contrast to fixed factors that are manipulated by the experimenter. In psycholinguistics these are usually Subject and Item.

Another advantage of mixed-effects modeling is that of *shrinkage*. By including the entire dataset, estimates for coefficients are inherently more precise than standard linear regression models (Baayen, 2008). This is achieved by estimating random effects for each level (i.e. Subject or Item) using the data of all available levels. The effect achieved thereby can be compared to that of regression towards the mean in repeated experiments. Therefore mixed-effects models offer a lot more statistical power than standard modeling techniques such as linear regression or ANOVA, especially in unbalanced experimental designs (Baayen, 2008; Kliegl, Wei, et al., 2010).

Model Fitting Strategy

A frequent approach of fitting linear mixed models is to use the maximal random effects (MRE) model structure (Barr, Levy, Scheepers, & Tily, 2013), i.e. keeping the random effects structure maximal with all possible random slopes, interactions and correlations. Bates, Kliegl, Vasishth, and Baayen (2015) disagree and argue that this approach is not useful for small datasets like they typically occur in behavioral research and could lead to overparameterization and thereby uninterpretable models. Bates, Kliegl, et al. (2015) offer a strategy to identify uninterpretable models and iteratively reduce complexity of the random effects structure to arrive at a *parsimonious model* for a given dataset.

The main diagnostic tool to identify uninterpretable models by checking dimensionality of the random effects structure is a principal components analysis (PCA). If that analysis discovers variance components to make a minimal contribution to the overall variance (less than 1%), then the model is found to be overparameterized and therefore uninterpretable (Bates, Kliegl, et al., 2015).

The first step as suggested by Bates, Kliegl, et al. (2015) is to check whether the PCA of the MRE model points to a degenerate model. In that case, the second step is to drop all correlation parameters which in turn leads to a zero correlation parameter (ZCP) model. If the PCA of the ZCP model points to overparameterization again (at least one component that contributes less than 1% of the overall variance), then in a third step one by one the smallest variance component is removed and once more the dimensionality of the resulting model is evaluated. This third step is repeated until the model's random effects structure is supported by its dimensionality.

By further dropping apparently negligible variance components, a model's fit can be improved with regard to the Akaike Information Criterion (AIC). An overparameterized model is usually a better fit to the data, but is degenerate and uninterpretable (Bates, Kliegl, et al., 2015). This is why a PCA should always precede and only models rendered interpretable should be compared.

Bates, Kliegl, et al. (2015) also suggest a fourth step in which correlation parameters are added to the model again. This step, however, is disregarded in the present analyses in order to translate the best fit zero correlation parameter LMMs to generalized additive models (GAMs)[1], which are the basis for the later trial history analyses.

Trial History Effects

The joint effect of quality and frequency is possibly not only modulated by the presence or absence of non-words. Masson and Kliegl (2013) found that in lexical decision tasks the interaction is modulated by lag effects (i.e. quality and frequency of the previous stimulus) and that an interaction could possibly be masked by trial history. Although particularly the former finding was criticized by other authors (Balota, Aschenbrenner, & Yap, 2013; O'Malley & Besner, 2013), it seems not far-fetched to consider that by aggregating data sets, the overall time course of an experiment could systematically conceal an interaction between quality and frequency in lexical processing. In addition to their finding that lag effects modulated that interaction, Masson and Kliegl (2013) also found a higher-order interaction with trial number

[1]Generally, a zero correlation parameter linear mixed model (LMM) is perfectly translatable to a generalized additive model (GAM) using the *mgcv* package in *R*. See General Discussion for remarks on the omission of correlation parameters in the presented approach.

when included as a covariate in the model. It is noteworthy that characteristics of previous trials, or the overall time course of an experiment itself (Bates, Kliegl, et al., 2015; Taylor & Lupker, 2001), could modulate or introduce so much unaccounted variance that small interactions get lost behind the noise. Modeling trial history effects, if there are any, might not only help to find possibly significant main or interactive effects, but be actually necessary in order to ensure the commonly required identical and independent distribution of the data (Bates, Kliegl, et al., 2015; Bates, Maechler, Bolker, & Walker, 2015).

Bates, Kliegl, et al. (2015) show that autocorrelated errors can be efficiently accounted for by yet another mixed-effects modeling technique, namely generalized additive models (GAMs, Wood, 2006), a subclass of generalized linear models. This allows to include trial number as wiggly fixed or random effects rather than assuming a linear function. Autocorrelation might be due to many different reasons, most obviously fatigue, attentional fluctuation or learning effects (Bates, Kliegl, et al., 2015; Taylor & Lupker, 2001), all of which are disadvantageous for statistical analyses. Usually we assume that those disappear by aggregating large data sets within and between subjects but if there is a systematic trend overall and/or within smaller experimental units such as blocks, data aggregation will hardly be helpful to eliminate those errors. Autocorrelation occurs when the value at a given point k correlates with the value at $t - k$, $k \geq 1$. If an autocorrelation plot shows a systematic trend, then clearly data aggregation alone was not helpful enough to account for those time-related effects.

In the light of past research, the following analyses are to investigate the additive, interactive and trial history effects of frequency and quality in naming experiments, both with and without non-words in the stimulus list. To examine the effects of trial history in naming tasks, two experiments were conducted in 2013 at the University of Victoria in Canada. Whereas the first one contained both words and non-words as trials, only words were tested in the second experiment. In addition to an examination of main and interactive effects in each experiment, this setup additionally allows a comparison of the experimental condition absence vs. presence of non-words.

With regard to trial history effects, all final LMMs are translated to GAMs as described above. The nature of the added fixed and random effects is to be guided by the ostensible shape

and constitution of a trial history plot (response times plotted over trial number). If the inclusion of trial history in the model improves the goodness of fit (regarding deviance explained and AIC), then the history-sensitive model is deemed advantageous over the history-naïve model. If differences in the pattern of significant effects are discovered, this is not a criterion of reasonableness of the presented approach, but rather a side effect of the approach that should be carefully evaluated. It is hypothesized that history-sensitive models improve the goodness of fit and clear up additional variance that can potentially mask small effects. A proposal for investigating trial history effects based on and partially including the Bates, Kliegl, et al. (2015) approach is covered and applied hereinafter.

Experiment 1

Introduction

The two experiments were designed to examine the effects of quality and frequency on reaction times in the presence or absence of non-words in the item list. Low-frequency and degraded stimuli were expected to exert slower responses. Frequency and quality were expected to exhibit additive effects for all contrasts as typically reported in the literature in the presence of non-words in the item list (Besner et al., 2010; Bonin et al., 2012; Carello et al., 1995; O'Malley & Besner, 2008). For this analysis, non-words were adopted as stimuli with zero frequency, i.e. they conform to the English grammar but are expected to be unknown to the reader. In fact, the experiment was designed to keep the participant unaware of the lexical status (word or non-word) of the stimulus presented.

As several authors suggested, trial history is examined in the present analyses since it is expected to affect response times under different conditions (Masson & Kliegl, 2013; Taylor & Lupker, 2001). The central and rather explorative question is whether the inclusion of trial history in the model helps reveal masked interactions, improves the fit or helps solve transformation issues.

Method

Subjects. A total of 72 participants were recruited from psychology students at the University of Victoria (Victoria, British Columbia, Canada) in 2013 and tested for this experiment to earn extra credit in an undergraduate psychology course.

Materials. High-frequency and low-frequency words (see Appendix B) were 4 to 5 characters long and taken from the *English Lexicon Project* database[2] (Balota et al., 2007). Non-words were matched to words using the *Wuggy* application[3] (Keuleers & Brysbaert, 2010) based on length, orthographic neighborhood and summed bigram frequency. All of the non-words followed English orthography and none of them were pseudohomophones. There were 120 high-frequency, 120 low-frequency and 240 non-word trials. As every stimulus was presented to each subject exactly once, in order to avoid possibly confounding effects of quality and Item, stimulus assignment to conditions was counterbalanced across subjects: Each stimulus was clear for one half of the subjects while being degraded for the other half. Degraded stimuli were 65% white and clear stimuli were black (0% white) on a white screen.

Procedure. The experiment was run using *SuperLab* on MacPro computers. Subjects were seated in front of the screen wearing a headset microphone to record the vocal response, informed about the procedure and materials, and asked to respond to all following stimuli as quickly and accurately as possible. For every subject 32 practice trials (50% non-word, 25% high-frequency and 25% low-frequency trials) and 480 critical trials were conducted. At the beginning of each trial a fixation cross (+) was shown in the center of the screen for 250 ms, followed by a blank screen for another 250 ms and the target stimulus in uppercase letters in the center of the screen until the subject made a vocal response. An experimenter scored the response as either correct or error using a keyboard. Trials that were unsuitable for further statistical analyses (such as extraneous noises before the actual pronounced stimulus) were marked as spoils and later removed from the dataset.

[2]The English Lexicon Project is available at http://elexicon.wustl.edu/

[3]The Wuggy application is available at http://crr.ugent.be/programs-data/wuggy

Results

All following analyses and plots were done using *R* (R Core Team, 2014) with packages *ggplot2* (Wickham, 2009), *plyr* (Wickham, 2011), *lme4* (Bates et al., 2015), *MASS* (Venables & Ripley, 2002), *RePsychLing* (Baayen, Bates, Kliegl, & Vasishth, 2015), *itsadug* (Rij, Wieling, Baayen, & Rijn, 2015), *lattice* (Sarkar, 2008) and *mgcv* (Wood, 2003, 2006, 2011). For *R* implementations of relevant models and calculations see Appendix C. As it can be expected that the nature of each item has a strong influence on the trial, mixed-model analyses were performed in order to account for both Item and Subject variance at once. The model fitting strategy described in the General Introduction section was applied.

Dependent measure. Sternberg (1969) postulated that when two factors are additive (non-interactive) on reaction time, then the underlying cognitive process might run in stages. Ever since then, one of the most popular dependent measures in cognitive psychology has been reaction time (RT). However, absence of evidence is not evidence of absence, so if two factors are found to be additive, this might be due to processing in stages, but it could as well be due to a masked interaction that is revealed only when using a more appropriate measure, experimental design and/or analytical instrument.

A preliminary Box-Cox power transformation check of the dataset returns $\lambda = -1.15$ which most closely corresponds to a reciprocal transformation on Tukey's Ladder of Power Transformations (Box & Cox, 1964; Tukey, 1977; Venables & Ripley, 2002). Reciprocal reaction time ($RRT = -RT^{-1}$), or *speed*, has already been discussed as a noteworthy alternative for RT by several authors in the past (e.g., Kinoshita, Mozer, & Forster, 2011; Kliegl, Masson, & Richter, 2010; Masson & Kliegl, 2013; Wainer, 1977) and more recently employed in analyses of Masson and Kliegl (2013). By reciprocal transformation the unit of RRT is Hertz ($1\text{Hz} = 1\text{s}^{-1}$), a metric that is related to other absolutely valid measures in well-established disciplines, such as rate in neuroscience or speed in physics. Some authors are skeptical towards transforming RT because it might systematically produce more underadditive interactions (Balota et al., 2013). Alternative procedures are briefly discussed in the General Discussion section.

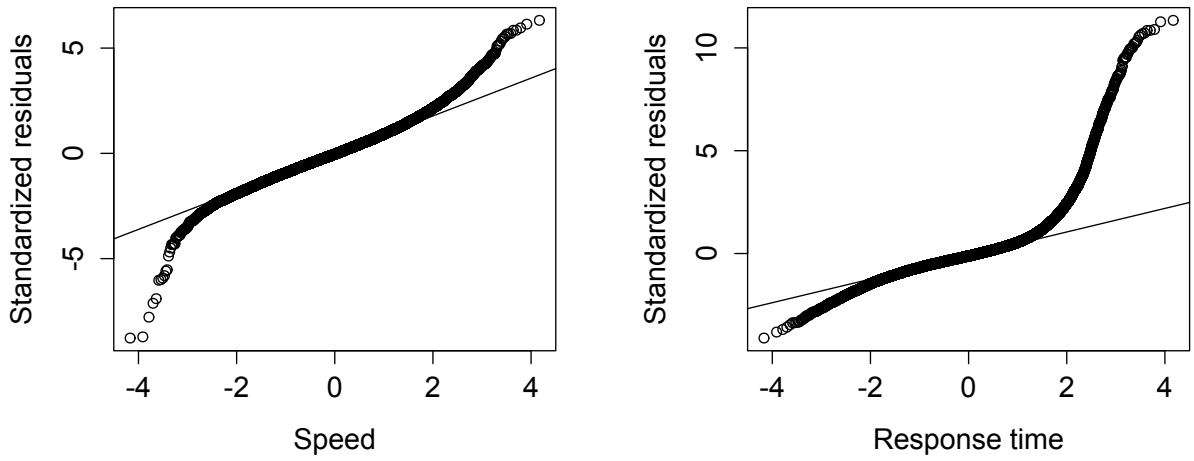

Figure 1. Q-Q plot of linear mixed-model residuals distributions using either RRT (left) or untransformed RT (right) as the dependent measure. The line passes through the 1st and 3rd quartiles.

Moreover, while raw RT may meet the requirements for many established analyses (e.g. ANOVA) it is not suitable for mixed-effects modeling using Gaussian distributions because residuals are usually not normally distributed but skewed (Kliegl, Masson, & Richter, 2010). Residuals of RRT mixed models, however, have been found to fit a normal distribution relatively well in lexical decision tasks (Kinoshita et al., 2011; Kliegl, Masson, & Richter, 2010; Masson & Kliegl, 2013). Thus the same transformation was used for the present analyses and found to fit the residuals to a normal distribution better than untransformed RT for naming as well (see Figure 1).

Mixed-model structure. The 2 × 3 experimental design yields a grand mean (intercept) and the two fixed factors quality with levels clear/degraded and frequency with levels non-word/low-frequency/high-frequency. The fixed factors and their respective joint effects were encoded as successive difference contrasts[4] (Venables & Ripley, 2002), generating a degraded–clear contrast for quality and a low-frequency–high-frequency and non-word–low-frequency contrast for frequency. The interactions of the quality contrast with both frequency contrasts (two interactions in total) were also included.

[4]Successive difference contrasts are specifically suited for ordered factors and do not require equal spacing between levels.

Table 1
Linear mixed-model variances and standard deviations for reciprocal RT (RRT) random effects in Experiment 1

Random effects	Variance	*SD*
Items		
Intercept	0.008	0.089
Quality	0.001	0.038
Subjects		
Intercept	0.044	0.210
Frequency		
LF-HF	0.001	0.037
NW-LF	0.003	0.055
Quality (D-C)	0.008	0.090
Frequency × Quality		
(LF-HF) × (D-C)	0.001	0.025
(NW-LF) × (D-C)	0.001	0.031
Residuals	0.048	0.219

LF-HF = low-frequency–high-frequency contrast; NW-LF = non-word–low-frequency contrast; D-C = degraded–clear contrast; SD = standard deviation

Random effects results. The MRE model comprised two random factors: Item and Subject. For the random factor Item there were intercept, the variance component for the within-item effect quality and a parameter for the correlation of intercept and quality. The random factor Subject included intercept, variance components for the within-subject effects of quality, both frequency contrasts and their respective interactions (one with quality per frequency contrast), as well as correlation parameters for all possible correlations between intercept and variance components (15 in total). Also taking into account the residual variance, there were 25 variance components and correlation parameters in the MRE model random effects structure. Expectedly, a PCA rendered this model uninterpretable: For Subject five out of six dimensions accounted for 100% of the variance explained, rendering the model degenerate.

Removing all correlation parameters from the initial MRE model leads to a ZCP model whose dimensions support all of the initial variance components. This suggests that none of the variance components need to be dropped from the ZCP model so that it is supported by the data. The random effects structure described above is kept for the final LMM in the present analysis with the exception of correlation parameters and contains a total of nine variance components, including residual variance (see Table 1).

Table 2
Linear mixed-model estimates of coefficients, standard errors, and t-values and generalized additive model p-values for RRT fixed effects in Experiment 1

Fixed effects	Coefficient	*SE*	*t*	*p*
Intercept	−1.428	0.025	−56.348	< 0.001
Frequency				
LF-HF	0.052	0.013	4.069	< 0.001
NW-LF	0.118	0.012	9.573	< 0.001
Quality (D-C)	0.170	0.011	15.191	< 0.001
Frequency × Quality				
(LF-HF) × (D-C)	0.011	0.009	1.251	0.211
(NW-LF) × (D-C)	−0.040	0.008	−4.924	< 0.001

LF-HF = low-frequency–high-frequency contrast; NW-LF = non-word–low-frequency contrast; D-C = degraded–clear contrast; SE = standard error

Table 3
Linear mixed-model estimates of coefficients, standard errors, and t-values and generalized additive mixed model p-values for (untransformed) RT fixed effects in Experiment 1

Fixed effects	Coefficient	*SE*	*t*	*p*
Intercept	753.067	16.719	45.044	< 0.001
Frequency				
LF-HF	33.046	10.047	3.289	< 0.002
NW-LF	85.640	11.731	7.300	< 0.001
Quality (D-C)	98.586	8.173	12.063	< 0.001
Frequency × Quality				
(LF-HF) × (D-C)	17.982	7.175	2.506	0.012
(NW-LF) × (D-C)	−5.095	8.613	−0.592	0.554

LF-HF = low-frequency–high-frequency contrast; NW-LF = non-word–low-frequency contrast; D-C = degraded–clear contrast; SE = standard error

Fixed effects results. Table 2 shows the estimated coefficients along with standard errors, *t*-values and significance levels for fixed effects in Experiment 1. As anticipated, main effects for quality ($t = 15.191$, $p < 0.001$) and frequency ($t = 4.069$, $p < 0.001$ for LF-HF and $t = 9.573$, $p < 0.001$ for NW-LF) are significant. Trials were slower for degraded words and lower word frequency. Interestingly, there seems to be an interaction between quality and frequency, specifically between the non-word–low-frequency contrast and quality for RRT: The quality effect was smaller for non-words than for low-frequency words (see Figure 2, $t = -4.924$, $p < 0.001$).

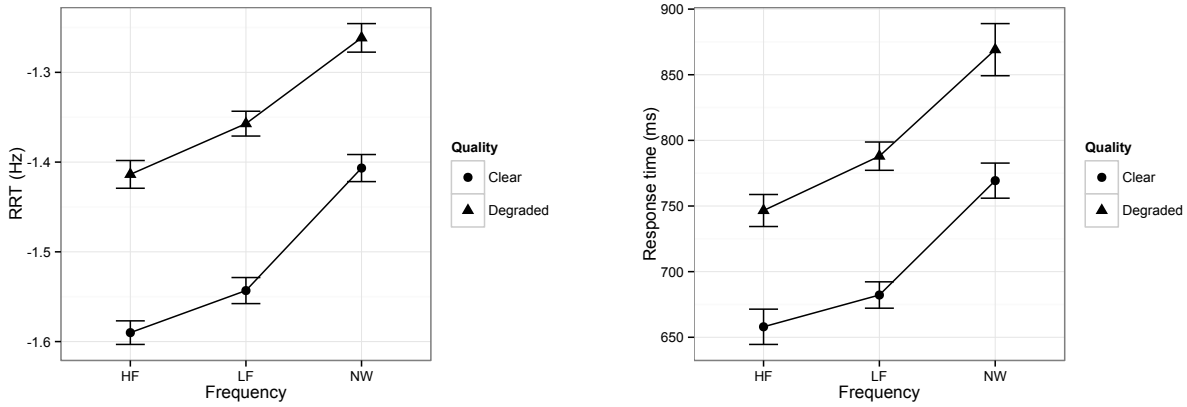

Figure 2. Mean reciprocal reaction time (RRT) and response time (RT) as a function of frequency and quality. Error bars are 95% within-subject confidence intervals appropriate for comparing condition means within a specific quality condition
HF = high frequency, LF = low frequency, NW = non-word

As Balota et al. (2013) criticized that reciprocal RT transformations might artificially create significant effects for some interactions, the data were fitted to another model using untransformed RT as the dependent variable. Results for untransformed RT fixed effects are shown in Table 3. Apparently, as visible in Figure 2, the underadditive effect from the RRT analysis is no longer significant ($t = -4.924, p < 0.001$ for RRT vs. $t = -0.592, p = 0.554$ for RT) but the previously additive joint effect is overadditive for RT ($t = 1.251, p = 0.211$ for RRT vs. $t = 2.506, p = 0.012$ for RT) so that the quality effect is significantly smaller for low-frequency than for high-frequency words. This was not an anticipated result. All other effects remain significant.

Trial history analysis. As described earlier, autocorrelated errors can have a substantial impact on the response. Especially in experimental settings such as the present one with a large number of relatively similar trials (specifically 480 trials words and non-words to be read aloud), it makes sense to include an account for autocorrelation in the model because learning effects and fatigue are to be expected and indeed seem to be present in the data (see Figure 3). The final zero correlation parameter LMM from above was used as the basis for fitting the data to a generalized additive model (GAM) in order to account for trial history by adding wiggly fixed and random effects. As suggested by Figure 3, there are two superimposed trial history curves: one recurring curve that is relatively similar between blocks and one overall trend towards slower responses. The fixed effect spline was incorporated as a thin-plate regression full tensor

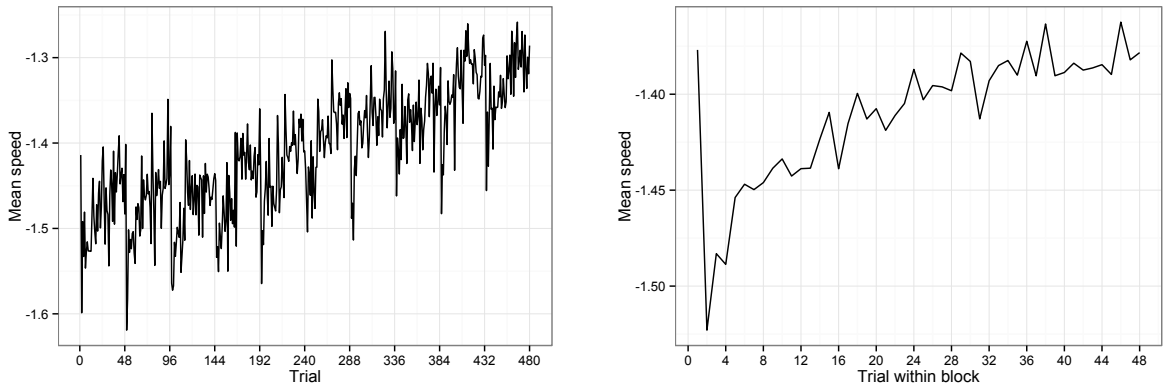

Figure 3. Mean speed (RRT) over trial history aggregated over subjects (left) and over subjects and blocks (right)

product smooth with 1st derivatives (Wood, 2003, 2006). The tensor product was preferred over a regular smooth to account not only for the block and overall within-block trend but also for a possible interaction between block and within-block trial number because it seems plausible that the within-block trend differs between blocks. Random splines were added as by-subject factor smooths with thin-plate regression and 1st derivatives for within-block trial and for block. It appeared neither possible nor plausible to model their interaction by Subject because every participant was only tested once for each trial number (i.e. once for each unique combination of block and within-block trial); the interaction is only between- but not within-subject and therefore only a fixed but not a random effect.

The approach of including trial history splines in the model proved appropriate since all smooth terms were estimated to be significant (ps < 0.001). Moreover, the resulting history-sensitive model ($df = 1213.34, AIC = -9681.44$) explains 65.1% of the deviance, while the earlier history-naïve model ($df = 886.28, AIC = -5563.83$) explains just 59.7%, because it was able to remove most of the autocorrelation from the residuals (see Figure 4). Furthermore, the shape of the fixed smoothing splines (see Figure A1) approximates the shape of the actual within-block mean speed (see Figure 3) extraordinarily well and composite splines of the wiggly fixed and random effects per subject offer insight into each participant's attentional fluctuation and fatigue over trials (see Figure A3). Although being a significantly better fit, neither coefficients nor t-values differ significantly from the previous model (compare Table 2 and 4).

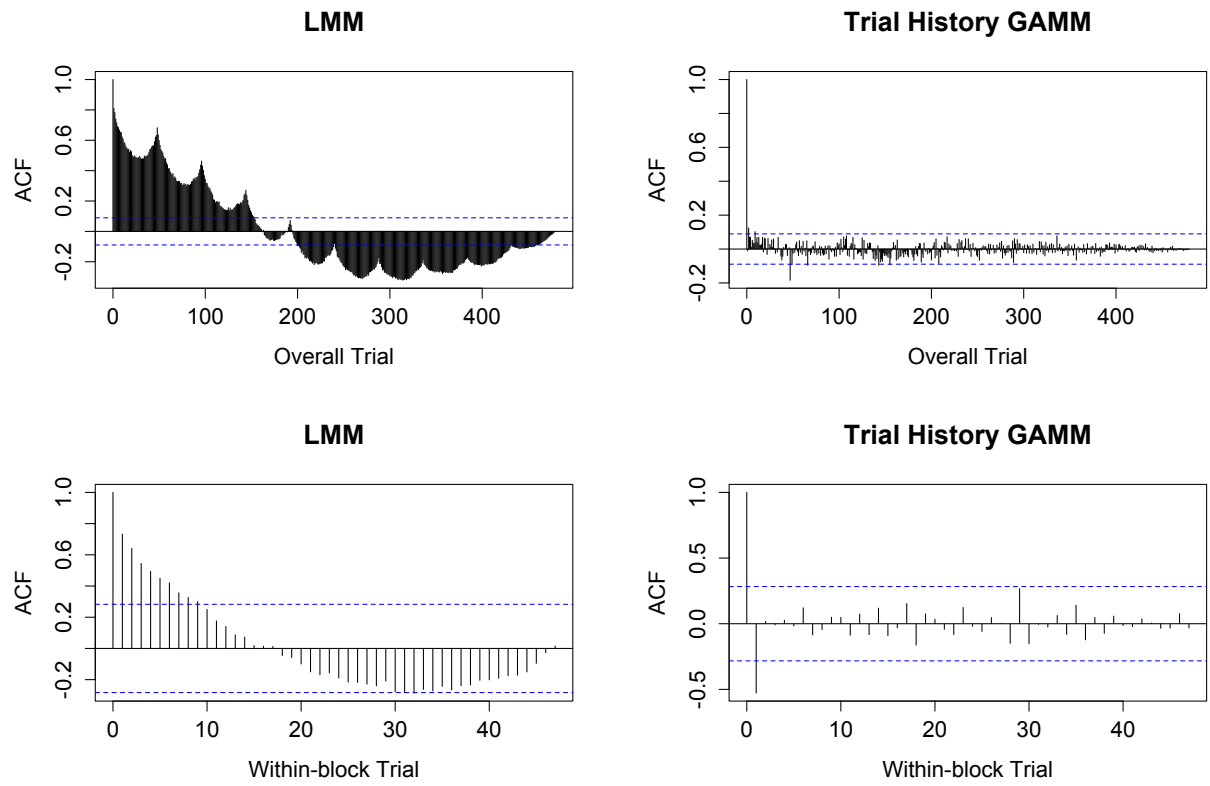

Figure 4. Autocorrelation function (ACF) plots for speed residuals by overall trial (top row) and within-block trial number (bottom row), for the history-naïve LMM (left) and the history-sensitive GAM (right)
Dashed lines are 95% confidence intervals

Table 4
Trial history GAM estimates of coefficients, standard errors, t-values and p-values for RRT fixed effects in Experiment 1

Fixed effects	Coefficient	*SE*	*t*	*p*
Intercept	−1.422	0.026	−55.429	< 0.001
Frequency				
LF-HF	0.052	0.013	4.042	< 0.001
NW-LF	0.117	0.012	9.501	< 0.001
Quality (D-C)	0.170	0.011	15.223	< 0.001
Frequency × Quality				
(LF-HF) × (D-C)	0.009	0.009	1.091	0.275
(NW-LF) × (D-C)	−0.041	0.008	−5.023	< 0.001

LF-HF = low-frequency–high-frequency contrast; NW-LF = non-word–low-frequency contrast; D-C = degraded–clear contrast; SE = standard error

Discussion

It was not surprising that frequency and quality had significant main effects on response time given present literature. Regarding the joint effects, however, there are some points that should be highlighted. Firstly, the interaction between the non-word–low-frequency contrast and quality was previously found to be additive in naming experiments with non-words by other authors (Besner et al., 2010; Bonin et al., 2012; Carello et al., 1995; O'Malley & Besner, 2008). In the present experiment, the effect of quality was smaller for non-words than for low-frequency words. Between low-frequency and high-frequency words, there was no significant difference in magnitude of the quality effect, which in turn is congruent with previous findings. This suggests that stimulus quality affects lexical processing more for known words than for unknown words.

Interestingly though, the pattern is somewhat different when analyzing untransformed RT. In that case, the joint effect between the non-word–low-frequency contrast and quality was additive. The low-frequency–high-frequency contrast, however, interacted with quality so that high-frequency words were less affected by stimulus quality than low-frequency words. So whereas RRT analyses produce standard findings with regard to the low-frequency–high-frequency by quality interaction, RT analyses show unexpected results. This is an interesting example illustrating the necessity for a resolution of the dependent measure debate. Under the assumption that RRT has better compliance with requirements, such as normal residual distribution, results of the RT analyses should be treated with caution. A post-hoc analysis showed that this result also holds with an ANOVA ($p < 0.01$), which is why the finding cannot only be due to a violation of the distribution requirements by untransformed RT.

Both the RRT and the RT result for the frequency-by-quality interaction suggest that frequency and quality are not purely additive, not even in the presence of non-words. There are some differences between the present experiment and previous studies that found additive effects. Firstly, although the onset of the stimulus was varied between previous studies, none seem to have used a 250 + 250 ms (250 ms fixation cross + 250 ms blank screen) delay. A longer delay between trials could imply more time for the word processor to be cleared. As

Ziegler, Perry, and Zorzi (2009) had already argued that the activation of the lexical route could critically depend on individual differences between readers, it seems plausible that the ratio of activation between lexical and nonlexical route is reset between trials or blocks, sufficient time given. An interaction between frequency and quality, as typically found in naming with words only, could have emerged for that reason in the present case as well.

It was hypothesized that including trial history effects in the model could possibly reveal interactive joint effects by accounting for some time-related variability (caused for example by attention fluctuation, learning effects or fatigue). For Experiment 1, this was not the case although the history-sensitive model was indeed a better fit to the data and did explain more deviance than the history-naïve model. Nonetheless, that is not an argument for ignoring trial history but rather supporting its inclusion. As apparent in Figure 3, there was a trend between and within blocks but for this experiment, effects created by experimental manipulation were simply strong enough to be detected without an account for trial history.

Experiment 2

Introduction

As the interaction of frequency and quality had been found to be modulated by the absence or presence of non-words in the item list, another experiment without non-words was conducted. Similarly to Experiment 1, low-frequency and degraded stimuli were expected to lead to slower responses. In contrast to the first experiment, however, in Experiment 2 an interaction between frequency and quality is expected as suggested by the literature (Besner et al., 2010; Bonin et al., 2012; Carello et al., 1995). Moreover, for this experiment, which was based on the experimental design of Experiment 1 but conducted independently, it is to examine whether there are similar between- and within-block trial-history effects and if they mask or produce an interaction between frequency and quality.

Method

Subjects. 76 psychology students from the University of Victoria (Victoria, BC, Canada) participated in 2013 to earn extra credit in an undergraduate psychology course.

Materials. Word items (low-frequency and high-frequency words) were identical to the words used for Experiment 1. There were no non-words. That amounts to 120 high-frequency and 120 low-frequency trials. The same item counterbalancing technique as for Experiment 1 was used: Each stimulus was clear for one half of the subjects while being degraded for the other half. Degraded stimuli were 65% white and clear stimuli were black (0% white) on a white screen and all stimuli were randomly intermixed.

Procedure. Hardware, software and instructions were identical to those for Experiment 1 except that subjects were only informed they would have to read aloud words (as there were no non-words in this design). For every subject 16 practice trials (half high-frequency and half low-frequency trials) and 240 critical trials were conducted. The rest of the experimental procedure was also identical to Experiment 1. Trials that were unsuitable for further statistical analyses (such as extraneous noises before the actual pronounced stimulus) were marked as spoils and later removed from the dataset.

Results

For Experiment 2, the same analytical tools and model fitting strategy as for Experiment 1 were used. For *R* implementations of relevant models see Appendix D.

Dependent measure. A Box-Cox power transformation check was performed for the dataset and returned $\lambda = -1.19$ which also approximates a reciprocal transformation on Tukey's Ladder of Power Transformations (Box & Cox, 1964; Venables & Ripley, 2002). Thus it seemed feasible to use RRT as the proper RT transformation for Experiment 2 as well.

Mixed-model structure. The 2 × 2 experimental design affords an intercept and the two fixed factors quality with levels clear/degraded and frequency with levels low-frequency/high-frequency. Both fixed factors were encoded as successive difference contrasts (Venables & Ripley, 2002), generating a degraded–clear contrast for quality and a low-frequency–high-frequency contrast for frequency. The interaction of quality and frequency was also included in the model.

Random effects results. The MRE model random effects structure consisted of two random factors: Item and Subject. For the random factor item there were intercept, the variance

Table 5
Linear mixed-model variances and standard deviations for reciprocal RT (RRT) random effects in Experiment 2

Random effects	Variance	*SD*
Items		
Intercept	0.006	0.075
Subjects		
Intercept	0.034	0.184
Frequency (LF-HF)	0.002	0.047
Quality (D-C)	0.006	0.080
Frequency (LF-HF) × Quality (D-C)	0.001	0.032
Residuals	0.041	0.204

LF-HF = low-frequency–high-frequency contrast; D-C = degraded–clear contrast; SD = standard deviation

component for the within-item effect quality and a correlation parameter for intercept and quality. The random factor Subject included intercept, variance components for the within-subject effects of quality, frequency and their interaction, as well as correlation parameters for all possible correlations between intercept and variance components (six in total). There were 14 variance components and correlation parameters in the MRE model random effects structure, including residuals. A PCA of the MRE model indicated overparameterization since some of the dimensions contributed less than 1% to the deviance explained by their respective random factor.

After removing all of the correlation parameters from the MRE model, the quality variance component for the random factor Item still seemed to contribute a negligible amount of the deviance explained and thus was dropped from the model. The final resulting mixed-model random effects structure is shown in Table 5. Including residuals, there were a total of six random intercepts and variance components in this model.

Fixed effects results. Coefficients, standard errors and *t*-values for this model are shown in Table 6. As for Experiment 1, main effects for quality ($t = 9.703, p < 0.001$) and frequency ($t = 6.824, p < 0.001$) are significant, but there is no evidence for an interaction between those. This is contradictory to the expectation. A trial history analysis is performed to examine whether that interaction is revealed after accounting for time-related fluctuation.

Table 6
Linear mixed-model estimates of coefficients, standard errors, and t-values and generalized additive model p-values for RRT fixed effects in Experiment 2

Fixed effects	Coefficient	*SE*	*t*	*p*
Intercept	−1.802	0.022	−83.170	< 0.001
Frequency (LF-HF)	0.078	0.011	6.824	< 0.001
Quality (D-C)	0.094	0.010	9.703	< 0.001
Frequency (LF-HF) × Quality (D-C)	0.010	0.007	1.404	0.160

LF-HF = low-frequency–high-frequency contrast; D-C = degraded–clear contrast; SE = standard error

Table 7
Generalized additive model estimates of coefficients, standard errors, t-values and p-values for RRT fixed effects in Experiment 2

Fixed effects	Coefficient	*SE*	*t*	*p*
Intercept	−1.803	0.021	−84.540	< 0.001
Frequency (LF-HF)	0.077	0.009	6.770	< 0.001
Quality (D-C)	0.093	0.011	9.828	< 0.001
Frequency (LF-HF) × Quality (D-C)	0.012	0.006	1.979	0.048

LF-HF = low-frequency–high-frequency contrast; D-C = degraded–clear contrast; SE = standard error

Trial history analysis. The parsimonious linear mixed model from above was translated to a GAM in order to include a fixed and by-subject random splines for trial history. The fixed effect spline was added as a full tensor product smooth with thin-plate regression and 1st derivatives and random splines were added as by-subject factor smooths with thin-plate regression and 1st derivatives for within-block trial and for block. In the history-sensitive model, the quality-by-frequency interaction variance component for Subject was turned non-significant and thusly dropped from the GAM specification. All remaining smooth terms, including the added wiggly fixed and random effects for trial number, were estimated to be significant (ps < 0.001).

Coefficients, standard deviations and t-values did not change significantly for intercept or main effects (see Table 7). However with the trial history splines included, the model was a better fit, explained more deviance (58.1%, $df = 710.39, AIC = -7222.91$) than the previous model (53.2%, $df = 441.86, AIC = -5802.66$) and could accordingly account for some variability that presumably masked the expected frequency-by-quality interaction before ($t = 1.979, p < 0.05$ vs. $t = 1.404, p = 0.160$ without trial history splines included). As visible

in Figure A2, the fixed effect smooths were able to capture the between-block trend (difference in spline means), and the typical within-block trend (spline average over blocks) plus by-block characteristics, notably similar to the effects captured in Experiment 1.

Fitting the same data using untransformed RT instead of RRT reveals the anticipated effects for quality ($t = 6.039, p < 0.001$) and frequency ($t = 6.950, p < 0.001$) as well as for their interaction ($t = 2.520, p < 0.012$). Including trial history in that model neither improves goodness of fit nor changes *t*-values or coefficients significantly.

Discussion

The mean response times for Experiment 2 were faster than for Experiment 1, supporting the idea that only the time-efficient lexical route is used for pronunciation. Analogous to Experiment 1 and as anticipated, frequency and quality exerted significant main effects on response time. Ignoring trial history in the model, the joint effect of quality and frequency was far from significant ($p = 0.160$). As opposed to the previous experiment, however, in this case the interaction was discovered to be significant ($p = 0.048$) after accounting for trial history. In fact, there was a notably high increase in the *t*-value of the joint effect. This interaction was actually anticipated by previous research but masked by the noise generated by time-related fluctuation. The discovered interaction revealed high-frequency words to be more affected by stimulus quality than low-frequency words, which is congruent with the expected result.

The RT analysis yielded the same results without an account for trial history. Including trial history splines in the RT model decreased the goodness of fit and is therefore not recommended for that particular case because it did not help explain more deviance. As residuals for this RT model were not normally distributed either, its results are again to be interpreted with caution.

General Discussion

As far as the account for trial history in mixed-model analyses of RRT is concerned, in both experiments the inclusion increased the goodness of fit. In one case it helped reveal an anticipated interactive effect that was just too small in magnitude and thusly hidden by the variance explainable by a spline for trial number. Generally speaking, including fixed and random

splines for trial history effects has been shown to be a valid way to explain more deviance and possibly detect small effects. If the data do not suggest a clear trend that could be captured with existing smoothing algorithms or if the resulting model does not improve the goodness of fit, then trial history should be ignored. However, in all remaining cases trial history should at least be briefly considered as a possible source of variance, especially in experiments where small effects are anticipated and time-related variability cannot be definitely ruled out. The time course of the experimental procedure has been shown to mask smaller effects. Critical evaluation should therefore be given to the experimental design.

One considerable drawback of the proposed procedure is truncating the fourth step of the Bates, Kliegl, et al. (2015) approach of model selection. Although in the presented cases models did not improve by including correlation parameters again, it is not unlikely that other models would. In order to include smoothing splines, however, correlation parameters have to be disregarded because the *mgcv* package does not support estimation of those (Wood, 2006, 2011). If correlation parameters are strictly needed, the *gamm4* package (Wood & Scheipl, 2014) should be used instead. Fitting models with *gamm4* allows estimation of correlation parameters but is not as statistically robust as *mgcv* and should only be used if correlation parameters *and* smoothing splines are needed for statistical assumptions or theoretical purposes.

In the present analyses, the interaction between frequency and quality was of particular interest. Depending on the dependent measure used in Experiment 1, quality interacted with either the low-frequency–high-frequency contrast or the non-word–low-frequency contrast. Neither was expected given the present literature. On the contrary, frequency and quality are expected to be additive if non-words are present in the item list. The same interaction was also found in Experiment 2, where it was, however, anticipated. Particularly interesting is that the estimated magnitude of the joint effect is relatively similar between the two experiments (see Table 4 and 7) but the variance of that particular effect in Experiment 1 is too large to safely distinguish it from zero. As results of RRT analyses differed from RT analyses especially with regard to significant interactions in Experiment 1, it is unclear if trial history is able to help resolve or avoid RT transformation issues that had previously been brought to the attention of researchers.

Garcia-Marques, Garcia-Marques, and Brauer (2014) recently pointed out that two-way interactions, even if statistically significant, are ambiguous and thusly uninterpretable if their lines do not cross over in an interaction plot. This is because these interactive patterns might be due to their main effects and a non-linear relationship between the latent and observed outcome, or due to an unmeasured mediator that has a non-linear relationship with the outcome variable. Actually, "[t]he smallest effect in a 2 × 2 ANOVA is always uninterpretable" (Garcia-Marques et al., 2014), be it main or interaction effect, and the same caution is to be taken with mixed-effects modeling. None of the aforementioned joint effects cross over which is why one should refrain from interpreting those as actual interactions. In the light of longer delays between items, however, the reader's word processor might be (partially) reset between trials and reactivate the lexical route, allowing effects similar to those found in naming experiments with words only. Even if the result is tentative, future research could address this question by selecting more contrasting material or vary between-trial delays in order to elicit cross-over interactions.

Using mixed-effects modeling is a relatively new technique and the establishment of guidelines about their proper use is an ongoing process. Several authors recently raised doubts about transforming response times for different reasons (e.g., Balota et al., 2013; Lo & Andrews, 2015; O'Malley & Besner, 2013). Most importantly, reciprocally transformed RT might systematically create more underadditive interactions than untransformed data. Lo and Andrews (2015) proposed to use inverse Gaussian generalized linear mixed models with identity link and untransformed RT instead of the herein applied Gaussian linear and generalized additive models with identity link and reciprocal RT. Although RRT fits residuals better to a normal distribution than any other known transformation (Lo & Andrews, 2015; Masson & Kliegl, 2013), inverse Gaussian fitting of untransformed RT was demonstrated to be a considerably good fit as well and, more importantly, allows direct inferences about mental chronometry. These inferences are difficult to make when using RRT due to a lack of a theoretical rationale to transform RT. Lo and Andrews (2015) argue that the selection of the DV should be guided by the research question rather than merely by mathematical assumptions. Nevertheless, this discussion is per se independent of the herein demonstrated advantage of trial history analyses since their potential applies to any dependent variable that is recorded over time.

Altogether trial history in all its facets is not to be ignored in statistical analyses. Manifested as lag effects, fatigue, recovery or attentional fluctuation, serial effects and time can impact the outcome of statistical analyses. Future research should address the question of how one can capture that source of variance in an efficient manner and how it could be incorporated in computational models such as CDP+. Taking all evidence into account, at least for now it seems evident that trial history is not history yet.

References

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

Baayen, R. H., Bates, D., Kliegl, R., & Vasishth, S. (2015). RePsychLing: Data sets from Psychology and Linguistics experiments. Retrieved from https://github.com/dmbates/RePsychLing

Balota, D. A., Aschenbrenner, A. J., & Yap, M. J. (2013). Additive effects of word frequency and stimulus quality: the influence of trial history and data transformations. *Journal of experimental psychology. Learning, memory, and cognition*, *39*(5), 1563–1571. doi:10.1037/a0032186.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., … Treiman, R. (2007). The English Lexicon Project. *Behavior research methods*, *39*, 445–459.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3). doi:10.1016/j.jml.2012.11.001

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious Mixed Models. Retrieved from http://arxiv.org/pdf/1506.04967

Bates, D., Maechler, M., Bolker, B. M., & Walker, S. (2015). lme4: Linear mixed-effects models using Eigen and S4. Retrieved from http://CRAN.R-project.org/package=lme4

Bates, D., Maechler, M., Bolker, B. M., & Walker, S. (in press). Fitting Linear Mixed-Effects Models using lme4. *Journal of Statistical Software*. Retrieved from http://arxiv.org/abs/1406.5823

Besner, D., O'Malley, S., & Robidoux, S. (2010). On the joint effects of stimulus quality, regularity, and lexicality when reading aloud: new challenges. *Journal of experimental psychology. Learning, memory, and cognition*, *36*(3), 750–764. doi:10.1037/a0019178

Bonin, P., Roux, S., Barry, C., & Canell, L. (2012). Evidence for a limited-cascading account of written word naming. *Journal of experimental psychology. Learning, memory, and cognition*, *38*(6), 1741–1758. doi:10.1037/a0028471

Borowsky, R. & Besner, D. (1993). Visual word recognition: A multistage activation model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(4), 813–840. doi:10.1037/0278-7393.19.4.813

Borowsky, R. & Besner, D. (2006). Parallel distributed processing and lexical-semantic effects in visual word recognition: are a few stages necessary? *Psychological Review*, *113*(1), 181–195. doi:10.1037/0033-295X.113.1.181

Box, G. E. P. & Cox, D. R. (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society B*, *26*, 211–252.

Carello, C., Lukatela, G., Peter, M., & Turvey, M. T. (1995). Effects of association, frequency, and stimulus quality on naming words in the presence or absence of pseudowords. *Memory & Cognition*, *23*(3), 289–300. doi:10.3758/BF03197231

Garcia-Marques, L., Garcia-Marques, T., & Brauer, M. (2014). Buy three but get only two: the smallest effect in a 2 × 2 ANOVA is always uninterpretable. *Psychonomic Bulletin & Review*, *21*(6), 1415–1430. doi:10.3758/s13423-014-0640-3

Keuleers, E. & Brysbaert, M. (2010). Wuggy: a multilingual pseudoword generator. *Behavior research methods*, *42*(3), 627–633. doi:10.3758/BRM.42.3.627

Kinoshita, S., Mozer, M. C., & Forster, K. I. (2011). Dynamic adaptation to history of trial difficulty explains the effect of congruency proportion on masked priming. *Journal of experimental psychology. General*, *140*(4), 622–636. doi:10.1037/a0024230

Kliegl, R., Masson, M. E. J., & Richter, E. M. (2010). A linear mixed model analysis of masked repetition priming. *Visual Cognition*, *18*(5), 655–681. doi:10.1080/13506280902986058

Kliegl, R., Wei, P., Dambacher, M., Yan, M., & Zhou, X. (2010). Experimental Effects and Individual Differences in Linear Mixed Models: Estimating the Relationship between Spa-

tial, Object, and Attraction Effects in Visual Attention. *Frontiers in psychology*, *1*, 238. doi:10.3389/fpsyg.2010.00238

Lo, S. & Andrews, S. (2015). To transform or not to transform: using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, *6*. doi:10.3389/fpsyg.2015.01171

Masson, M. E. J. & Kliegl, R. (2013). Modulation of additive and interactive effects in lexical decision by trial history. *Journal of experimental psychology. Learning, memory, and cognition*, *39*(3), 898–914. doi:10.1037/a0029180

O'Malley, S. & Besner, D. (2008). Reading aloud: qualitative differences in the relation between stimulus quality and word frequency as a function of context. *Journal of experimental psychology. Learning, memory, and cognition*, *34*(6), 1400–1411. doi:10.1037/a0013084

O'Malley, S. & Besner, D. (2013). Reading aloud: does previous trial history modulate the joint effects of stimulus quality and word frequency? *Journal of experimental psychology. Learning, memory, and cognition*, *39*(4), 1321–1325. doi:10.1037/a0031673

Perry, C., Ziegler, J. C., & Zorzi, M. (2007). Nested incremental modeling in the development of computational theories: the CDP+ model of reading aloud. *Psychological Review*, *114*(2), 273–315. doi:10.1037/0033-295X.114.2.273

R Core Team. (2014). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from http://www.R-project.org/

Rij, J. v., Wieling, M., Baayen, R. H., & Rijn, H. v. (2015). itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs.

Sarkar, D. (2008). *Lattice: Multivariate data visualization with R*. Use R! New York: Springer Science+Business Media.

Stanners, R. F., Jastrzembski, J. E., & Westbrook, A. (1975). Frequency and visual quality in a word-nonword classification task. *Journal of Verbal Learning and Verbal Behavior*, *14*(3), 259–264. doi:10.1016/S0022-5371(75)80069-7

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica*, *30*, 276–315. doi:10.1016/0001-6918(69)90055-9

Taylor, T. E. & Lupker, S. J. (2001). Sequential effects in naming: A time-criterion account. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(1), 117–138. doi:10.1037//0278-7393.27.1.117

Treisman, A. M. & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136. doi:10.1016/0010-0285(80)90005-5

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Venables, W. N. & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed). Statistics and computing. New York: Springer.

Wainer, H. (1977). Speed vs reaction time as a measure of cognitive performance. *Memory & Cognition*, *5*(2), 278–280. doi:10.3758/BF03197375

Wickham, H. (2009). *Ggplot2: elegant graphics for data analysis*. Springer New York. Retrieved from http://had.co.nz/ggplot2/book

Wickham, H. (2011). The Split-Apply-Combine Strategy for Data Analysis. *Journal of Statistical Software*, *40*(1), 1–29. Retrieved from http://www.jstatsoft.org/v40/i01/

Wood, S. N. (2003). Thin-plate regression splines. *Journal of the Royal Statistical Society (B)*, *65*(1), 95–114.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.

Wood, S. 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, S. N. & Scheipl, F. (2014). gamm4: Generalized additive mixed models using mgcv and lme4. Retrieved from http://CRAN.R-project.org/package=gamm4

Yap, M. J. & Balota, D. A. (2007). Additive and interactive effects on response time distributions in visual word recognition. *Journal of experimental psychology. Learning, memory, and cognition*, *33*(2), 274–296. doi:10.1037/0278-7393.33.2.274

Ziegler, J. C., Perry, C., & Zorzi, M. (2009). Additive and interactive effects of stimulus degradation: no challenge for CDP+. *Journal of experimental psychology. Learning, memory, and cognition*, *35*(1), 306–311. doi:10.1037/a0013738

Appendix A

Spline figures

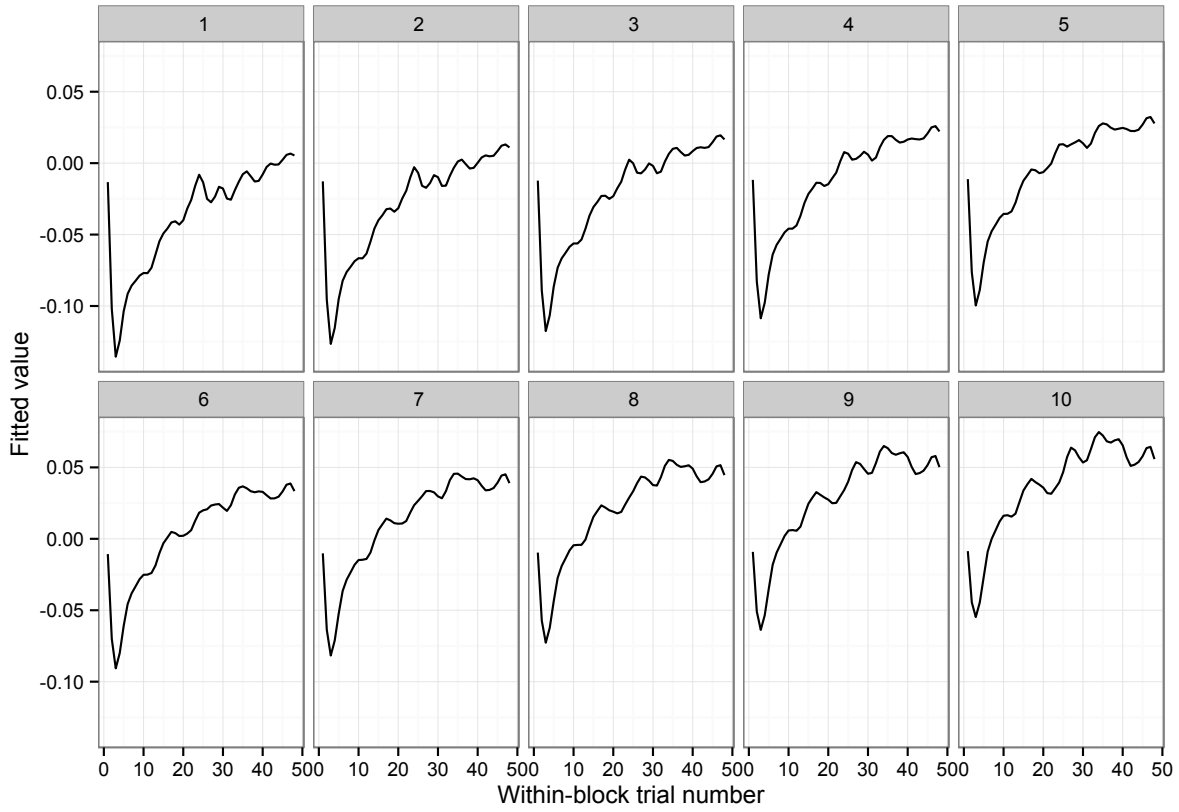

Figure A1. Wiggly fixed effect splines of within-trial number and block number in Experiment 1

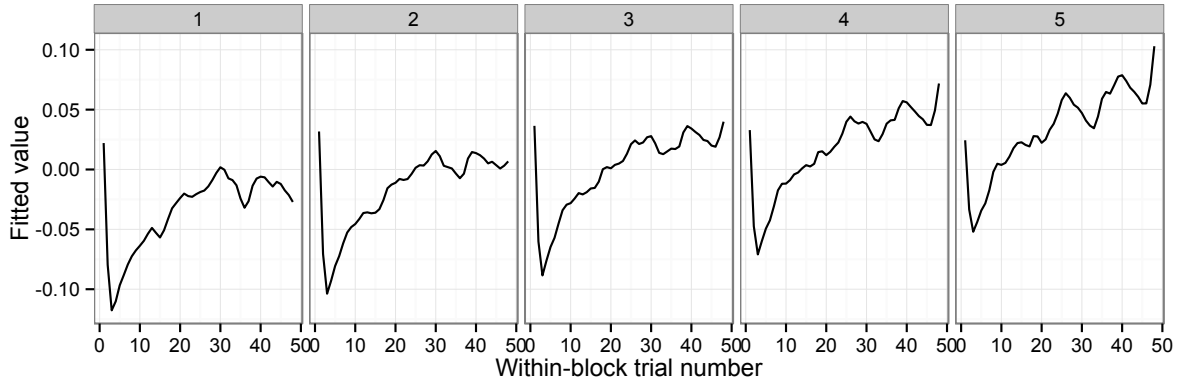

Figure A2. Wiggly fixed effect splines of within-trial number and block number in Experiment 2

(Appendix continues)

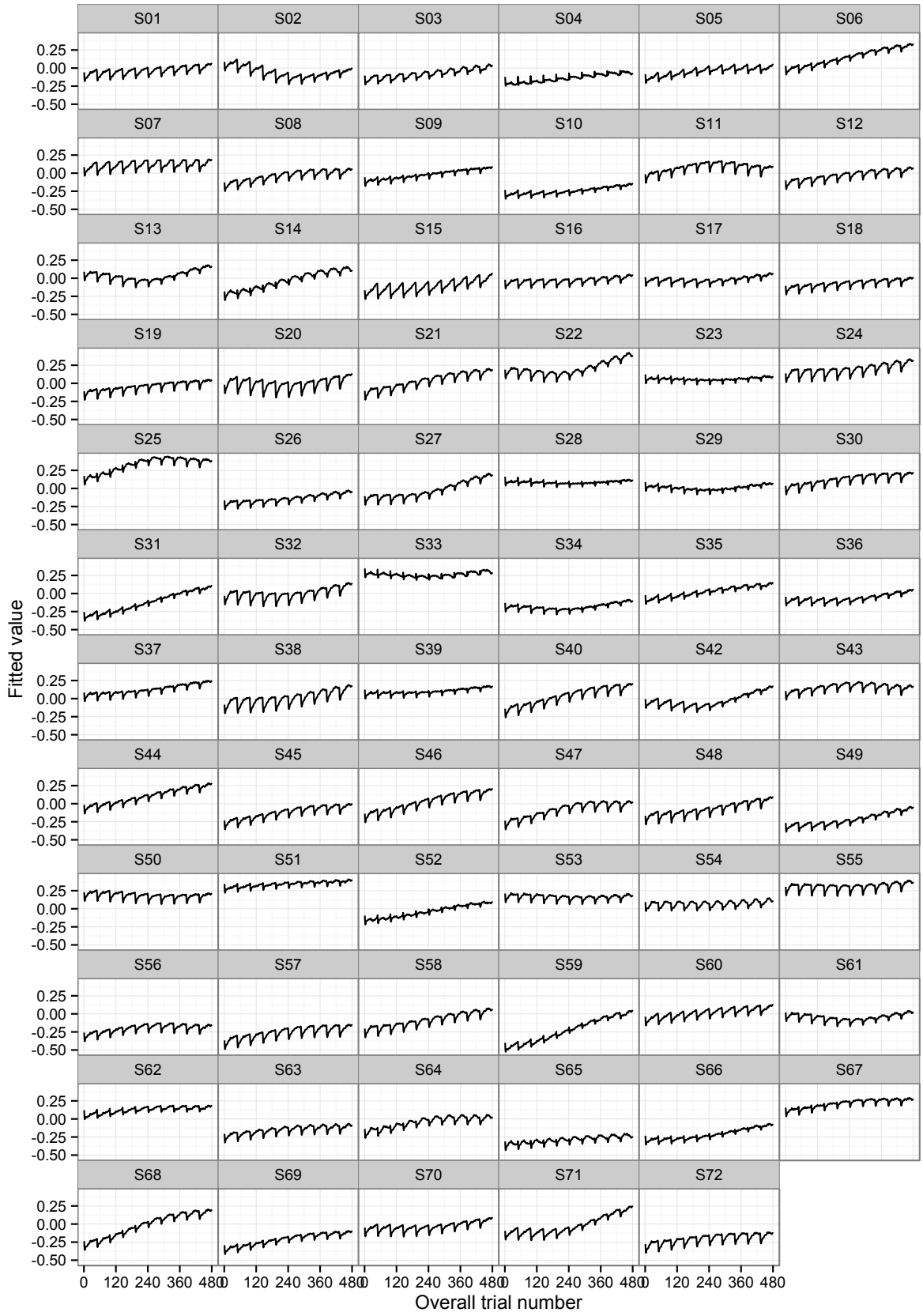

Figure A3. Composite wiggly fixed and random effect splines as a function of overall trial number and subject in Experiment 1

Appendix B

Stimuli

High-frequency

PLAN	EARTH	AFTER	FORCE	CLUB	RULE
MAJOR	YEAR	CLASS	EXIST	MUSIC	WOMAN
WROTE	QUITE	ONCE	WANT	ONLY	TODAY
NEAR	DRIVE	REPLY	LAND	FAST	PLACE
PARTY	BLACK	TOTAL	NOTE	PIECE	COST
LIVE	READ	MEAN	SEND	NICE	FIRST
STOP	BAND	WORLD	CHILD	SIDE	MOVIE
WHITE	HOUSE	MONEY	FREE	MODEL	HEART
ISSUE	SOON	ROUND	STAY	MOST	MAGIC
COURT	ALLOW	FIRE	DATA	CHECK	SELL
RACE	TRACK	FOUND	SPACE	CARD	FALL
SMALL	EXTRA	LARGE	POST	STUDY	TABLE
LEAVE	CALL	LEAD	HAND	WORTH	VOTE
COLOR	PART	ORDER	EARLY	JUST	BASIC
BUILD	GUESS	SOUND	TALK	SPEED	HUMAN
APPLE	TEAM	LEGAL	WATCH	TEXT	YOUNG
SHORT	PAST	LEVEL	BREAK	SURE	THANK
NORTH	LOCAL	WHEN	TITLE	CAUSE	IDEA
EVEN	NEED	WALL	LAST	OPEN	ABLE
OFFER	GROUP	WHOLE	AGREE	TYPE	MANY

Low-frequency

STUB	LOFT	PRAWN	CARVE	NIECE	SPECK
SKUNK	ULCER	CHORE	CLOUT	SHALE	SCARF
TIDY	SOOT	HUNCH	GROPE	TULIP	JINX
PLUM	GLOAT	CRAB	LADLE	SNARL	PEST
BRISK	MANGO	TINT	CRAVE	CRAMP	SPREE
SCALD	TRAMP	MUGGY	EMBER	POUCH	DUSK
PERCH	KNEEL	HIKER	SNOB	PLANK	FERN
JEER	SWAT	PERK	HAZE	TOTEM	FANG
BROOM	LICE	HOVER	SKIM	WEEP	GERM
OATS	STRUT	APRON	POISE	CHUG	MEND
TONIC	THAW	PETAL	STRUM	BRAN	TEMPT
SCOLD	HEED	HOARD	MISTY	BOAST	GRIME
BRAT	DECOY	FAWN	SMOG	TART	HEDGE
GLARE	PANDA	FLIRT	SLUM	VEAL	TROT
TIMID	GUST	MOLAR	CARP	FROWN	LARK
FABLE	HONK	SLEEK	SEAM	TWINE	PEAR
OMEN	SNEER	SNORT	REGAL	WAKEN	SCORN
SLOB	MINK	STORK	TANGO	RAKE	SLEET
OTTER	BROTH	BRAID	HAUNT	VALET	BINGE
TWIG	WAFER	SNORE	GREET	ASPEN	FLANK

(Appendix continues)

Non-words

SNAN	AXLER	CLEB	CRUB	PRATH	NIELS
MAHUR	CLACE	RUSAC	GLUNK	CHONE	SHAND
TWOTE	OVED	IPLY	FIMY	HUTHS	TUMEP
NEAK	DETLY	FAWS	SNUM	CREB	SNALK
SARBY	HOMAL	PIEZE	CLISK	SINT	CRAND
MIVE	MEAS	NIDE	SCARD	ROGGY	POUTH
STOM	WODDS	SIRT	PEXED	HULER	PLAVE
WHINT	SONEM	SEDEL	ZEER	PEWN	HOTEW
ISTUB	ROUTH	MONG	SHOOM	HUKER	WOAP
ROURT	FIME	CHEGS	OAMS	ASCAN	SPUG
RARD	DOUND	CAZE	HONIX	DOTAL	BRON
SMOLL	LALKS	STEVY	SCOLT	HOAFS	BOUST
MEAVE	TEAD	WORBS	TRAT	FAWL	TADE
COSIR	UMDER	JUSH	GLASH	FLORT	JEAL
BUITS	SOUNT	PREED	HIDID	SOTAR	SCOWN
ALQUE	TEFAL	TEPH	TADLE	FLEEK	TWIND
SHOYS	LODEL	SUSH	OTAN	SNURT	GAYEN
NORLS	SWEN	CAIFS	CROB	STOIN	RASS
EGAN	WAMS	IWEN	UCTER	CRAID	VUMET
UDFER	WHOSS	MYPE	TWEG	SNOSH	AGSEN
EARNT	FONTH	RUCH	LORT	CARCE	CLECK
CEAR	ELINS	GOLAN	UTFER	PROUT	SCARB
QUIMP	WABS	TOGEY	DOOT	CROPE	JITT
PLIVE	LART	PLART	WHOAT	FATLE	PEBS
BLADS	NOKE	CORS	PANVO	CRANG	SPROU
REAN	SEFF	FIRPS	SPAMP	ESSER	DULK
BAFF	CHIZE	MEDIE	TWEEL	THOB	FERT
HOUTH	BLEE	MEART	SWAD	HAMB	FAND
HOON	STAW	RAWIC	LIRD	TWIM	GESK
ALLAB	RAMA	SECK	STRIT	PONTH	MERT
PRACK	SPAPE	RALL	THAB	STRUG	TELSH
ERSHA	PODE	FADLE	HEEN	MALKY	GRIMP
CANG	HAFF	VOMP	DEBOP	GLOG	HERKS
PAFT	EIFLY	CADIC	SANNA	TRUM	CROT
GURGH	TASP	HYLAN	GUNT	CARF	LART
TOOM	WASTS	YOUST	HOMS	HEAM	PEAM
PACH	FLEAK	THAVE	SNEEG	REWAT	BLORN
LOMEL	HIDRE	UVYA	MIGS	FANVO	CREET
JEED	LARE	ASHO	BRODE	HAIDS	BIDGE
GROOP	ATHIE	CALY	GACER	PREET	GRANK

Appendix C

R Implementation of Experiment 1

```
———————— Box-Cox Transformation Check ————————
lambdaList <- boxcox(rt ~ Subj*Q*T, data=d)
(lambda <- lambdaList$x[which.max(lambdaList$y)])

———————————— Final LMM ————————————
s1 <- lmer(speed ~ 1 + f + n + q + f_q + n_q + (1 + f + n + q +
      f_q + n_q || Subj) + (1 + q || Item), data=d, REML=FALSE)
print(summary(s1), corr=FALSE)

———————— Principal Components Analysis (PCA) ————————
(pca <- rePCA(s1))
pca$Item$sdev^2*100/sum(pca$Item$sdev^2)
pca$Subj$sdev^2*100/sum(pca$Subj$sdev^2)

———————— Final LMM Translated to History-Naive GAM ————————
g0 <- bam(speed ~ 1 + f + n + q + f_q + n_q +
      s(Subj, bs="re") + s(Subj, f, bs="re") +
      s(Subj, n, bs="re") + s(Subj, q, bs="re") +
      s(Subj, f_q, bs="re") + s(Subj, n_q, bs="re") +
      s(Item, bs="re") + s(Item, q, bs="re"),
      data=d, method="ML")
print(summary(g0))

———————————— History-Sensitive GAM ————————————
g1xnc <- bam(speed ~ 1 + f + n + q + f_q + n_q +
      te(block, btrial, m=1, k=c(3,39)) +
      s(btrial, Subj, bs="fs", k=5, m=1) +
      s(block, Subj, bs="fs", k=3, m=1) +
      s(Subj, bs="re") + s(Subj, f, bs="re") +
      s(Subj, n, bs="re") + s(Subj, q, bs="re") +
      s(Subj, f_q, bs="re") + s(Subj, n_q, bs="re") +
```

(Appendix continues)

```
23        s(Item, bs="re") + s(Item, q, bs="re"),
24        data=d, method="ML")
25  print(summary(g1xnc))
```

Appendix D

R Implementation of Experiment 2

The commands for the Box-Cox power transformation check and the PCA are identical to those for Experiment 1 (see ll. 1–2, 6–8).

```
                                  Final LMM
1   s2 <- lmer(speed ~ 1 + q*f + (1 + q*f || Subj) + (1 | Item),
2         data=d, REML=FALSE)
3   print(summary(s2), cor=FALSE)

            Final LMM Translated to History-Naive GAM
4   g0 <- bam(speed ~ 1 + q*f + s(Subj, bs="re") +
5         s(Subj, q, bs="re") + s(Subj, f, bs="re") +
6         s(Subj, q_f, bs="re") + s(Subj, bs="re"),
7         data=d, method="ML")
8   summary(g0)

                        History-Sensitive GAM
9   g2xnc <- bam(speed ~ 1 + q*f +
10        te(block, btrial, m=1, k=c(3,39)) +
11        s(btrial, Subj, bs="fs", k=5, m=1) +
12        s(block, Subj, bs="fs", k=3, m=1) +
13        s(Subj, bs="re") + s(Subj, q, bs="re") +
14        s(Subj, f, bs="re") + s(Item, bs="re"),
15        data=d, method="ML")
16  summary(g2xnc)
```

Appendix E

Statutory Declaration

I hereby declare that I completed this work on my own and that information which has been directly or indirectly taken from other sources has been noted as such. Neither this, nor a similar work, has been published or presented to an examination committee.

Potsdam, September 4, 2015	
Place and date	Maximilian Michael Rabe

Anhang F

Selbstständigkeitserklärung

Ich erkläre hiermit, dass ich diese Arbeit selbstständig angefertigt und sämtliche Informationen, die unmittelbar oder indirekt aus anderen Quellen entnommen wurden, als solche kenntlich gemacht habe. Weder diese noch eine ähnliche Arbeit wurde zuvor veröffentlicht oder einem Prüfungskomitee vorgestellt.

Potsdam, 4. September 2015	
Ort und Datum	Maximilian Michael Rabe