The Cave of Shadows
Addressing the human factor with generalized additive mixed models

Harald Baayen[a]

Shravan Vasishth[b]

Reinhold Kliegl[b]

Douglas Bates[c]

[a] University of Tübingen, Germany

[b] University of Potsdam, Germany

[c] University of Wisconsin-Madison, USA

Running Head: The Cave of Shadows

Corresponding author:

R. Harald Baayen

Seminar für Sprachwissenschaft

Eberhard Karls University Tübingen

Wilhelmstrasse 19

Tübingen

e-mail: harald.baayen@uni-tuebingen.de

Abstract

Generalized additive mixed models are introduced as an extension of the generalized linear mixed model which makes it possible to deal with temporal autocorrelational structure in experimental data. This autocorrelational structure is likely to be a consequence of learning, fatigue, or the ebb and flow of attention within an experiment (the 'human factor'). Unlike molecules or plots of barley, subjects in psycholinguistic experiments are intelligent beings that depend for their survival on constant adaptation to their environment, including the environment of an experiment. Three data sets illustrate that the human factor may interact with predictors of interest, both factorial and metric. We also show that, especially within the framework of the generalized additive model, in the nonlinear world, fitting maximally complex models that take every possible contingency into account is ill-advised as a modeling strategy. Alternative modeling strategies are discussed for both confirmative and exploratory data analysis.

Keywords: generalized additive mixed models, factor smooths, within-experiment adaptation, autocorrelation, experimental time series, confirmatory and exploratory data analysis, model selection

All models are wrong, but some are useful.
George Box (1979)

1 Introduction

Regression models are built on the assumption that the residual errors are identically and independently distributed. Mixed models make it possible to remove one source of non-independence in the errors by means of random-effect parameters. For instance, in an experiment with fast and slow subjects, the inclusion of by-participant random intercepts ensures that the fast subjects will not have residuals that will tend to be too large, and that the slow subjects will not have residuals that are too small (see, e.g. Pinheiro & Bates, 2000, for detailed examples). However, even after including random-effect parameters in a linear model, errors can still show non-independence.

For studies on memory and language, it has been known for nearly half a century that in time series of experimental trials, response variables such as reaction times elicited at time t may be correlated with earlier reaction times at $t - k, k \geq 1$ (Broadbent, 1971; Welford, 1980; Sanders, 1998; Taylor and Lupker, 2001; Baayen and Milin, 2010). One source of temporal dependencies between trials is the presence of an autocorrelational process in the errors, potentially representing fluctuations in attention. Another source may be habituation to the experiment, possibly in interaction with decisions made at preceding trials (Masson and Kliegl, 2013). Alternatively, subjects may slow down in the course of an experiment due to fatigue. A further source of correlational structure in sequences of responses is learning. As shown by Marsolek (2008), the association strengths between visual features and object names are subject to continuous updating. Ramscar, Yarlett, Dye, Denny and Thorpe (2010) and Arnon and Ramscar (2012) documented the consequences of within-experiment learning in the domain of language. Kleinschmidt and Jaeger (2015) report and model continuous updating in auditory processing in the context of speaker-listener adaptation. De Vaan, Schreuder and Baayen (2007) reported lexical decisions at trial t to be co-determined by the lexicality decision and the reaction time to a prime that occurred previously at $t - 40$. Grammaticality judgements that change in the course of an experiment are reported by Dery and Pearson (2015). In what follows, we refer to the ensemble of learning, familiarization with the task, fatigue, and attentional fluctuations as adaptive processes, or, in short, the 'human factor'. We also refer to data in which the human factor plays no role whatsoever as 'sterile' data, data that are not infected in any way by hidden processes unfolding in time series of experimental trials.

In what follows, we discuss three data sets that are demonstrably infected and not sterile. Why might we expect that experimental data are not sterile? Because, unlike molecules or plots of barley, human beings adapt quickly and continuously to their environment, and as the work mentioned above has shown, this includes the environment of psycholinguistic experiments.

When temporal autocorrelations are actually present in the data, but not brought into the statistical model, the residuals of this model will be autocorrelated in experimental time. The proper evaluation of model components by means of t or F tests presupposes that residual errors are identically and independently distributed. By bringing random intercepts and random slopes into the model specification, clustering in the residuals by item or subject is avoided. However, such random slopes and random intercepts do not take care of potential trial-to-trial autocorrelative structure. The presence of autocorrelation in the residuals leads to imprecision in model evaluations and uncertainty about the validity of any significances reported. When strong autocorrelation characterizes the residuals, this uncertainty will make it impossible to draw well-founded conclusions about statistical significance.

It might be argued that adaptive processes, if present, will have effects that are so minute that they are effectively undetectable. If so, the experimental design, and only the experimental design, could serve as a guide for determining the statistical model to be fitted to the data. Alternatively, one might acknowledge the presence of adaptive processes but claim that their presence gives rise to random and temporally uncorrelated noise. Any such adaptive processes would therefore be expected not to interact with predictors of theoretical interest.

However, it is conceivable that adaptive processes are present in a way that is actually not harmless. We distinguish two cases. First, adaptive processes may be present, without interacting with critical predictors of theoretical interest. In this case, measures for dealing with the autocorrelation in the errors will be required, without however affecting the interpretation of the predictors. In this case, elimination of autocorrelation from the errors will result in p-values that are more trustworthy. Second, it is in principle possible that adaptive processes actually do interact with predictors of theoretical interest in non-trivial ways. If so, it is not only a potential autocorrelational process in the residual error that needs to be addressed, but also and specifically the adaptive processes. These processes, which themselves may constitute a considerable source of autocorrelation in the errors, will need to be examined carefully in order to provide a proper assessment of how they modulate the effects of the critical predictors.

In what follows, we discuss three examples of non-sterile data demonstrably infected by adaptive processes unfolding in the experimental time series constituted by the successive experimental trials. First, we re-analyze a data set with multiple subjects, and a 2 × 2 × 4 factorial design with true treatments (Kliegl, Kuschela and Laubrock, 2015) and a single stimulus ‘item’. We then consider a mega-study with auditory lexical decision (Ernestus and Cutler, 2015) using a regression design with crossed random effects of subject and item. Finally, we investigate a self-paced reading study in which subjects were reading Dutch poems, following up on earlier analyses presented in Baayen and Milin (2010).

The analyses of these three data sets make use of the generalized additive mixed model (GAMM). Before presenting these analyses, we first provide a brief introduction to GAMMs. In the general discussion, we discuss strategies for dealing with the human factor when conducting confirmatory or exploratory data analysis.

2 The generalized additive mixed model

Linear regression models a univariate response y_i (where i the number of data points) as the sum of a linear predictor η and a random error term with zero mean. This linear predictor is assumed to depend on a set of predictor variables. Often, the response variable is assumed to have a normal distribution. If so, a regression model such as

$$y_i = \eta_i + \epsilon_i \text{ where } \epsilon_i \underset{\text{ind}}{\sim} N(0, \sigma^2) \text{ and } \eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}.$$

describes a response variable y that is modeled as a weighted sum of two predictors, x_1 and x_2, together with an intercept (β_0) and Gaussian error with standard deviation σ.

Generalized linear models let the response depend on a smooth monotonic function of the linear predictor. This family of models allows the response to follow not only the normal distribution, but other distributions from the exponential family, such as Poisson, gamma, or binomial. An example of a binomial GLM with the same linear predictor η is

$$y_i \underset{\text{ind}}{\sim} \text{binom}(\exp(\eta_i)/\{1 + \exp(\eta_i)\}, 1) \text{ where } \eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}.$$

Here, $\exp(\eta_i)/\{1 + \exp(\eta_i)$ is a reparameterization of the probability parameter in the binomial distribution, and 1 is the 'number of trials' (sample size). The generalized linear mixed model (GLMM) enriches the GLM with further sources of random noise, modeled with the help of Gaussian random variables with mean zero and unknown standard deviation to be estimated from the data. By way of example, if y denotes response time, x_1 the amount of sleep deprivation, and x_2 temperature, an experiment carried out with multiple subjects j would be analyzed with the model

$$y_{ij} = \eta_{ij} + \epsilon_{ij} \text{ where } \epsilon_{ij} \underset{\text{ind}}{\sim} N(0, \sigma^2) \text{ and } \eta_{ij} = \beta_0 + b_j + \beta_1 x_{1i} + \beta_2 x_{2i}, \text{ with } b_j \underset{\text{ind}}{\sim} N(0, \sigma_b^2),$$

under the assumption that the only term in the model that has to be adjusted from subject to subject is the intercept. In other words, this model assumes that there are faster and slower subjects, and that in all other respects, subjects behave in the same way. Specifically, the effects of the predictors x_1 and x_2 are assumed not to vary across subjects. More complex models can be obtained by relaxing these assumptions (see, e.g., Pinheiro & Bates, 2000). The b_j given the estimate of σ_b are known as best unbiased linear predictors (BLUPs), conditional modes, or posterior modes.

A generalized *additive* mixed model (Hastie & Tibshirani, 1990; Lin & Zhang, 1999; Wood, 2006, 2011, 2013; Wood, Goude & Shaw, 2015) is a GLMM in which part of the linear predictor η is itself specified a sum of smooth functions of one or more predictor variables. By way of example, consider a model for reaction time in the visual lexical decision task (transformed to approximately normal by the function $f(y) = -1000/y$) with frequency of occurrence as predictor.[1] Using the by-item mean RTs for the younger age group in the English Lexicon Project (Balota, Cortese, Sergent-Marshall, Spieler & Yap, 2004) as available for 2293 monomorphemic and monosyllabic words in the data set `english` in the `languageR` package (Baayen, 2010), we first consider a model with a linear effect of frequency (x):

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \text{ where } \epsilon_i \underset{\text{ind}}{\sim} N(0, \sigma^2).$$

The AIC for this model is -2953. The effect of frequency, however, is curvilinear, as already observed by Balota et al. (2004). We therefore need to relax the linearity assumption, and allow y to be a smooth function f of x:

$$y_i = \beta_0 + f(x_i) + \epsilon_i, \text{ where } \epsilon_i \underset{\text{ind}}{\sim} N(0, \sigma^2).$$

The smooth $f(x)$ is a weighted sum of a set of q so-called basis functions defined over x (James, Witten, Hastie & Tibshirani, 2013). Writing B_k for the k-th basis function, we have that

$$f(x_i) = \sum_{k=1}^{q} B_k(x_i) w_k.$$

When a cubic polynomial is used to model the nonlinear effect of x, the basis functions $B_k, k = 1, \ldots, 4$ are $B_1(x) = 1, B_2(x) = x, B_3(x) = x^2$, and $B_4(x) = x^3$. Polynomials, unfortunately, have various undesirable properties when interest is in the behavior of the response variable over the full range of the predictor. For instance, the left panel of Figure 1 shows the smooth for frequency when a polynomial of degree 4 is used (AIC: -3111). At the edges of the domain of frequency, where data are sparse, the polynomial shows wiggliness that in all likelihood is artifactual. The spline-based smooth in the right panel presents a more careful assessment of the nonlinear functional relation of reaction time and frequency (AIC: -3112).

[1] The transformation converts from time to speed metric; multiplication with -1 simply makes larger scores represent slower performance in the time metric (as is customary in psychology). In the absence of theoretical reasons for one of the metrics (i.e., hypotheses are at the ordinal level), time (mostly used in psychology) or speed (mostly used in physics) are both equally plausible. In the current context, the speed metric has the advantage of meeting assumptions of the preferred inferential statistics.

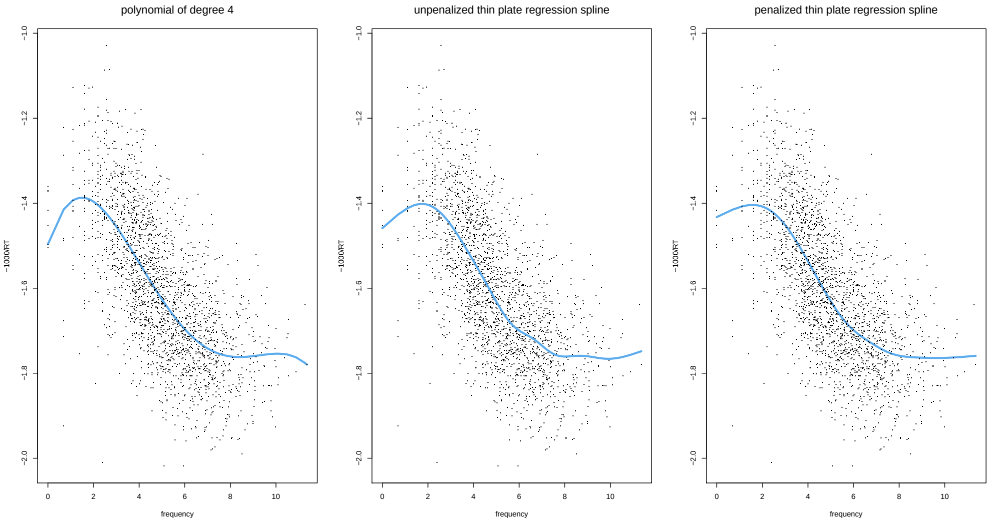

Figure 1: Reaction time (negative inverse transformed) as a smooth function of frequency of occurrence, with 2SE confidence intervals. Left panel: polynomial fit; center panel: thin plate regression spline without penalization for wiggliness; right panel: thin plate regression spline with penalization for wiggliness. Note that penalization results in a smoother regression curve.

Splines typically make use of basis functions B_k that have less extreme curvature than higher-order polynomials. Of the many smoothing splines that have been developed, the thin plate regression spline (TPRS) is the one that has the best performance with respect to minimizing the mean squared error. The basis functions of the thin plate regression splines are lower-order linearly independent polynomials. Figure 2 graphs the basis functions for the TPRS smooth for frequency. With the weights shown on the Y-axis, the smooth shown in the center panel of Figure 1 is obtained. These weights, however, slightly overfit. By imposing a penalty on wiggliness, the parameters (weights) for the basis functions can be chosen in such a way that an optimal balance is found between keeping the model simple (by giving less weight to highly wiggly basis functions) and staying faithful to the data. The (properly penalized) thin plate regression spline is shown in the right panel of Figure 1. It is more conservative at the edges of the frequency domain, and is smoother for the higher frequencies.

For a smooth based on polynomials, the number of polynomial basis functions (the degree of the polynomial) has to be selected, and the analyst has no guidance other than trying out which degree works best. For GAMMs, the number of basis functions is also part of the model specification. For the GAMM for frequency, the number of basis functions was set to 10. However, all that is happening here is that a number of basis functions is specified that is sufficiently large. The algorithm will

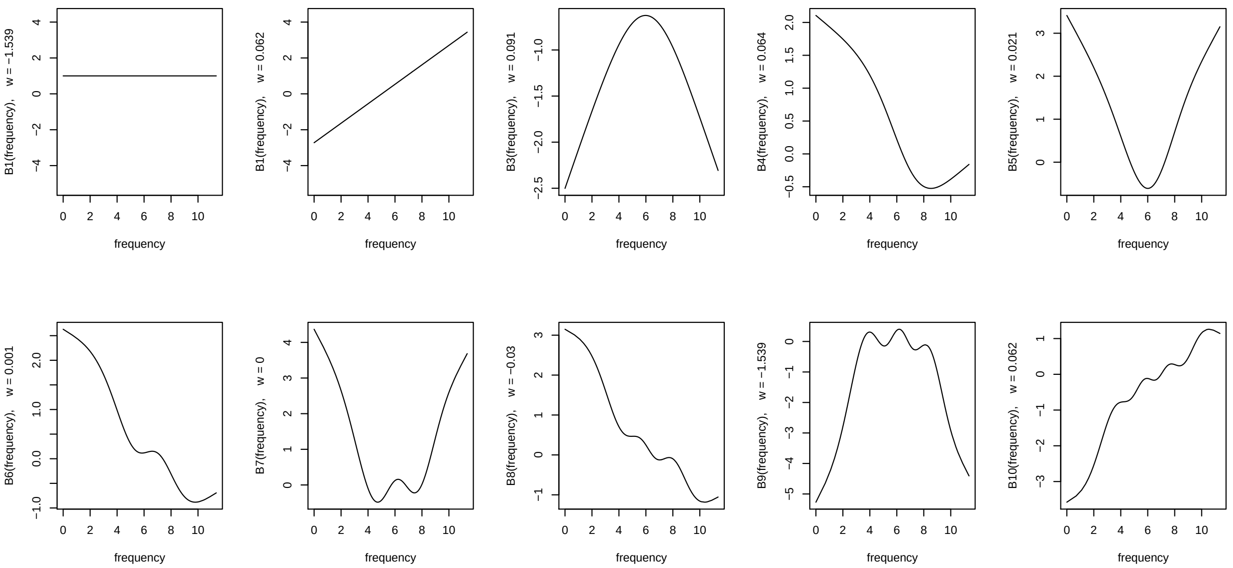

Figure 2: Example basis functions for a thin plate regression spline for `frequency`. On the vertical axis, the weights w are specified when there is no penalization for wiggliness. The fitted curve obtained by summation of these weighted basis functions is shown in the second panel of Figure 1.

then find the proper dimensionality given the data, which can be much lower than the upper bound specified by the user. Thus, a predictor that is strictly linear will be found to be linear by the algorithm, even when it is initially provided with a large number of basis functions. In such a case, the second basis function of Figure 2 suffices to capture the effect.

The effective degrees of freedom (edf) of a smooth specify the extent to which the freedom to vary when there is no penalization for wiggliness has been restricted by the penalization. For the smooth for frequency in the right panel of Figure 1, there are 10 basis functions, and hence we have, before penalization, 10 degrees of freedom, one for each of the corresponding weights. After penalization, the effective degrees of freedom are 6.37, indicating that the estimated curve is less wiggly than the curve estimated without penalization. If the frequency effect had been linear, then the effective degrees of freedom would have been reduced to 1: The linear basis function and its parameter would have been sufficient for modeling the effect. The edf for a smooth can informally be understood as the sum of the factors (between 0 and 1) by which each of the corresponding weight parameters have been shrunk due to penalization.

A Bayesian approach to variance estimation is employed (by putting a prior on wiggliness on the assumption that the truth is more likely to be smooth than wiggly). This makes for easier confidence interval calculation while at the same time providing good coverage probabilities (Nychka, 1988; Marra & Wood, 2012).[2]

Random effects are implemented as parametric terms penalized by a ridge penalty (James et al., 2013), which is equivalent to the assumption that the coefficients are independently and identically distributed normal random effects. The implementation of random effects by means of ridge penalties does not exploit the sparse structure of many random effects, and hence they are more costly to compute than corresponding random effects in the linear mixed model.

For the interaction of a metric predictor with a random effect factor, the linear mixed model offers the combination of random slopes and random intercepts. In this way, different regression lines (with shrinkage estimates) are obtained for each of the levels of the random effect factor. The factor smooth interaction of the generalized additive mixed model offers the possibility to relax the linearity assumption, and to fit wiggly curves for each of the factor levels. With appropriate penalties, they are fully 'random effects' and the nonlinear counterpart of the combination of random intercepts and random slopes in the linear mixed model.

A generalized additive (mixed) model is additive in two ways. First, it inherits from the generalized linear model that the linear predictor is a weighted sum. The generalized additive model adds to this functions of one or more predictors that themselves are weighted sums of basis functions. An important property of GAMMs is that each term in the model specifies a partial effect, i.e., the effect of that specific term when all other terms in the model are held constant. By way of example, Figure 3 presents a (simplified) model for visual lexical decision latencies for Vietnamese compound nouns (see Pham & Baayen, 2015, for a more comprehensive analysis) with the specification

$$y_i = \beta + \alpha_{t(i)} + f_1(x_{1i}) + f_2(x_{2i}, x_{3i}) + \epsilon_i, \text{ where } \epsilon_i \underset{\text{ind}}{\sim} N(0, \sigma^2),$$

where f_1 is a univariate smooth function, and f_2 a bivariate smooth interaction (technically represented using a 'tensor product' smoother). Let $t(i)$ denote which level of factor α observation i relates to. We can now let $\alpha_{t(i)}$ specify the effect of the tone realized on the first constituent of a Vietnamese compound being the most common (mid-level) tone, or some other tone (levels TRUE, FALSE, coded with treatment coding). Furthermore, let x_1 denote log word frequency, and x_2 and x_3 denote the frequencies of the first and second constituents.

[2] For a fully Bayesian approach to generalized additive modeling, see Wood (2016), where an interface between mgcv and the BUGS language is discussed. With this interface, it becomes possible to adopt a fully Bayesian approach to GAMMs.

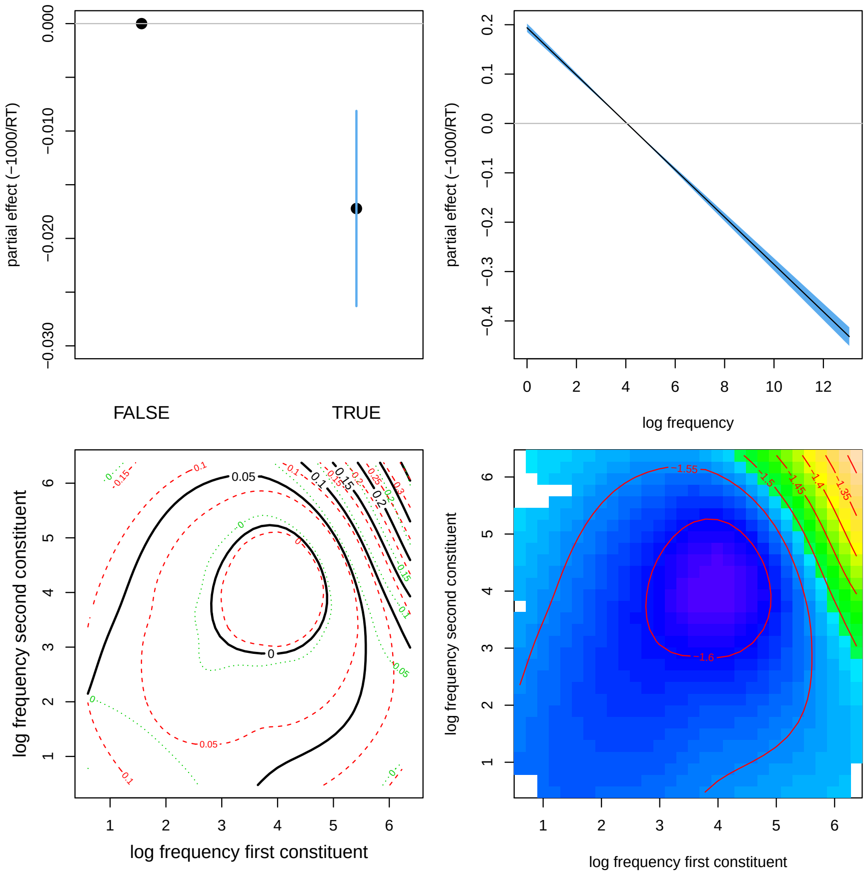

Figure 3: A simplified model for Vietnamese lexical decision latencies to compound nouns, with an effect of tone (upper left), word frequency (upper right), and the left and right constituent frequencies (bottom panels). The lower left panel shows contour lines with 1SE confidence intervals, the lower right panel presents the corresponding contour plot. Lighter and warmer colors denote longer response latencies.

Figure 3 shows the *partial effects*, i.e., the contributions of the individual terms in the model. For the factorial contrast in the upper left panel, the reference level is at zero, and we find that compounds with a mid-level tone are on average -0.0172 units faster on the -1000/RT scale. Frequency was entered into the model with a thin plate regression spline, but its effect is linear, and it is this linear effect that is returned by the smooth. The confidence intervals for the line have width zero where they intersect with the horizontal line crossing 0 on the y-axis. This is because for a GAM to be identifiable, all uncertainty about the intercept (where frequency is zero) is already quantified through the standard deviation for the intercept.

The bottom panels illustrate a nonlinear interaction, here involving the two constituent frequencies, modeled with a tensor product smooth. Shortest responses are found for intermediate values, the longest response times occur when both frequencies are high. The lower left panel shows contour lines with 1SE confidence intervals, red dashed lines represent the lower interval, and green dotted lines the higher interval. The contour plot in the lower right facilitates interpretation with color coding. Deeper shades of blue indicate shorter reaction times.

It is often the case that a covariate has a functional form that differs for the individual levels of a factor. In such cases, different wiggly curves can be fit to each factor level. Similarly, smooth regression surfaces can be allowed to vary across factor levels.

$$y_i = \beta + \alpha_{t(i)} + f_1(x_{1i}, \text{by} = t(i)) + f_2(x_{2i}, x_{3i}, \text{by} = t(i)) + \epsilon_{ij}, \text{ where } \epsilon_i \underset{\text{ind}}{\sim} N(0, \sigma^2),$$

where $f_i(x_1, x_2, \ldots, \text{by} = t(i))$ denotes the smooth for the interaction of $x_1, x_2, \ldots$ by α. It is important that for models with these kind of interactions, the main effect of the factor $(\alpha_{t(i)})$ is an important component of the model, as it has the crucial function of properly calibrating the different curves, surfaces (or hypersurfaces) with respect to the intercept.

The analyses in this study were carried out with the help of the `mgcv` package (Wood, 2006, 2011) and the `itsadug` package (van Rij, Baayen, Wieling and van Rijn, 2015) for `R` (R Development Core Team, 2015). These analyses are all exploratory, in that a sequence of increasingly complex models was constructed and only those predictors and interactions were maintained that received substantial support for improving the model fit.

3 The KKL dataset

The experiment reported by Kliegl et al. (2015), a follow up to Kliegl, Wei, Dambacher, Yan and Zhou (2011), showed that validly cued targets on a monitor are detected faster than invalidly cued ones, i.e., spatial cueing effect (Posner, 1980) and that targets presented at the opposite end of a rectangle at which the cue had occurred were detected faster than targets presented at a different rectangle but with the same physical distance, an object-based effect (Egly, Driver and Rafal, 1994). The sequence of an experimental trial is shown in Figure 4. Different from earlier research, the two rectangles were not only presented in cardinal orientation (i.e., in horizontal or vertical orientation), but also diagonally (45 degrees left or 45 degrees right). This manipulation afforded a follow up of a hypothesis that attention can be shifted faster diagonally across the screen than vertically or horizontally across the screen (Kliegl et al., 2011; Zhou, Chu, Li and Zhan, 2006). Finally, data are from two groups of subjects, one group had to detect small targets and the other large targets.

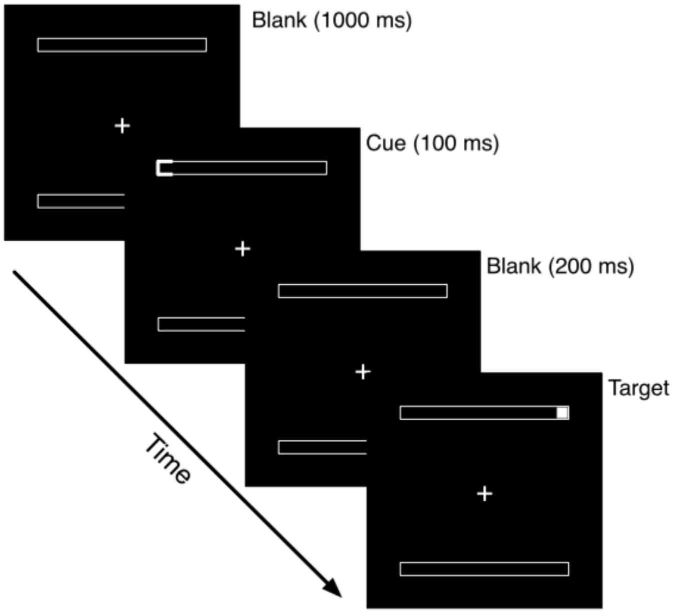

Figure 4: Sequence of events in visual-spatial attention experiment with an invalid cue on the same object. Screen 2: Left end of top rectangle is cued. Screen 3: SOA of 200 ms, Screen 4: Large target to be detected at right end of top rectangle (from Kliegl et al., 2011). In the new experiment rectangles were also presented in diagonal orientations; a different group of subjects was tested with small targets.

For an interpretation of fixed effects relating to the speed of visual attention shifts under these experimental conditions we refer to Kliegl et al. (2015).

Eighty-six subjects participated in this experiment. There were 800 trials requiring detection of a small or large rectangle and 40 catch trials. The experiment is based on a size (2) × cue-target relation (4) × orientation (2) design. Targets were small or large; rectangles were displayed either in cardinal or diagonal orientation, and cue-target relation was valid (70% of all trials) or invalid in three different ways (10% of trials in each the invalid conditions), corresponding to targets presented (a) on the same rectangle as the cue, but at the other end, (b) at the same physical distance as in (a), but on the other rectangle, or (c) at the other end of the other rectangle. Size of target was varied between subjects, the other two factors within subjects. The three contrasts for cue-target relation test differences in means between neighboring levels: spatial effect, object effect, and gravitation effect (Kliegl et al., 2011). Orientation and size factors are also included as numeric contrasts in such a way that the fixed effects estimate the difference between factor levels. With this specification the intercept estimates the grand mean of the 16 (= $2 \times 4 \times 2$) experimental conditions. The data are available as KKL in the `RePsychLing` package at https://github.com/dmbates/RePsychLing. Dependent variables is the log of reaction time for correct trials completed within a 750-ms deadline. The total number of responses was 53765.

Bates, Kliegl, Vasishth and Baayen (2015) determined a parsimonious mixed model for these data, dealing with issues of overparameterization. We refitted this model using in addition a quadratic polynomial, which allowed us to include a well-supported nonlinear effect for stimulus onset asynchrony, which was varied randomly in an interval ranging from 300 to 500 ms in this experiment, but the effect of which had not been included in the initial LMM report of Bates et al. (2015).

Model criticism is an important but all too often neglected part of data analysis. Inspection of the residuals of the reference model reveals that although the residuals approximately follow a normal distribution, and although they are identically distributed, they are not independent.

The reaction times of a given subject constitute a time series, with experimental trial as unit of time. These trials can be ordered from the initial trial in the experimental list of trials, to the final trial in that list. In what follows, we refer to the time series of trials with the covariate `Trial` $= 1, 2, \ldots, k$. For the present experiment, we have 86 such time series, one for each of the 86 subjects. When we consider the residuals of the reference model, ordered by these time-series, we observe autocorrelative structure.

The strength of the autocorrelations in these by-subject time series varied from subject to subject, as illustrated for four exemplary subjects in the top panels of Figure 5. The autocorrelations for the subject in the top left panel are quite mild, and unlikely to adversely affect model statistics. For the second subject, we find evidence for autocorrelations up to at least lag 3. Autocorrelations increase for the third subject, and are still present at a lag of 15 trials. The subject in the rightmost panel show strong autocorrelations, indicating that a response time at trial t is remarkably well correlated with the response at time $t - L$, for lags L as large as 25.

Given that the residuals of the reference model (refitted with a GAMM) are not independent, it is unclear how reliable the estimates of model parameters and the assessments of the uncertainty about these estimates actually are. For this particular data set, strong autocorrelations such as for the last subject are exceptional, and hence it is likely that conclusions based on this model will be somewhat accurate. Nevertheless, a statistical model that is formally deficient is unsatisfactory,

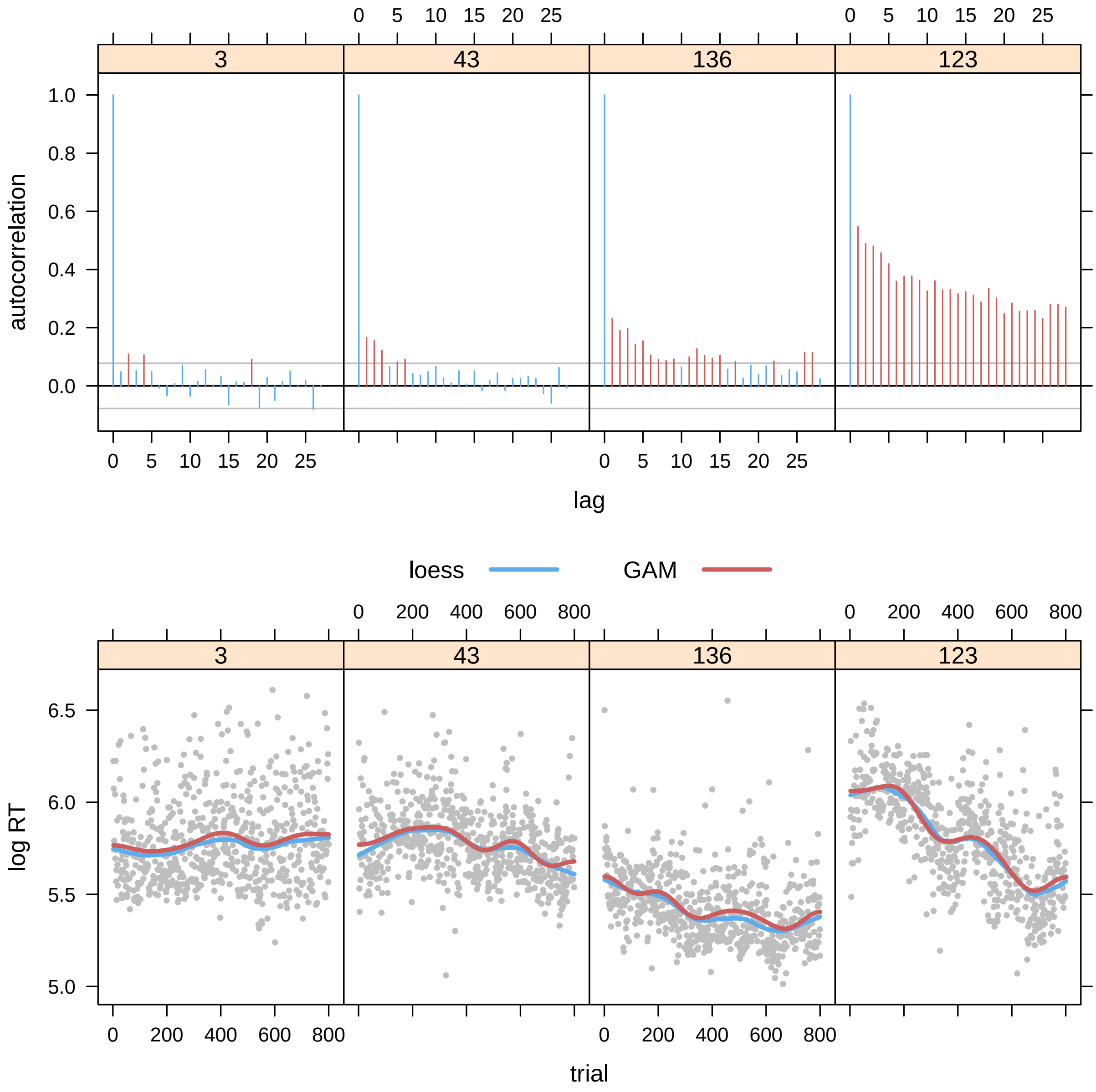

Figure 5: Autocorrelation functions for four subjects in the KKL dataset (upper panels), and the corresponding plots (lower panels) graphing log self-paced reading latency against trial, with a loess smoother (span = 0.2, in blue) and a GAM factor smooth (red).

especially as there must be hidden temporal processes unfolding in this experiment that are not transparent to the analyst. Since the KKL data are clearly not sterile, a more fertile approach is to bring such hidden processes out in the open, and incorporate them into the statistical model.

Why are these autocorrelations present? In order to address this question, consider the plots

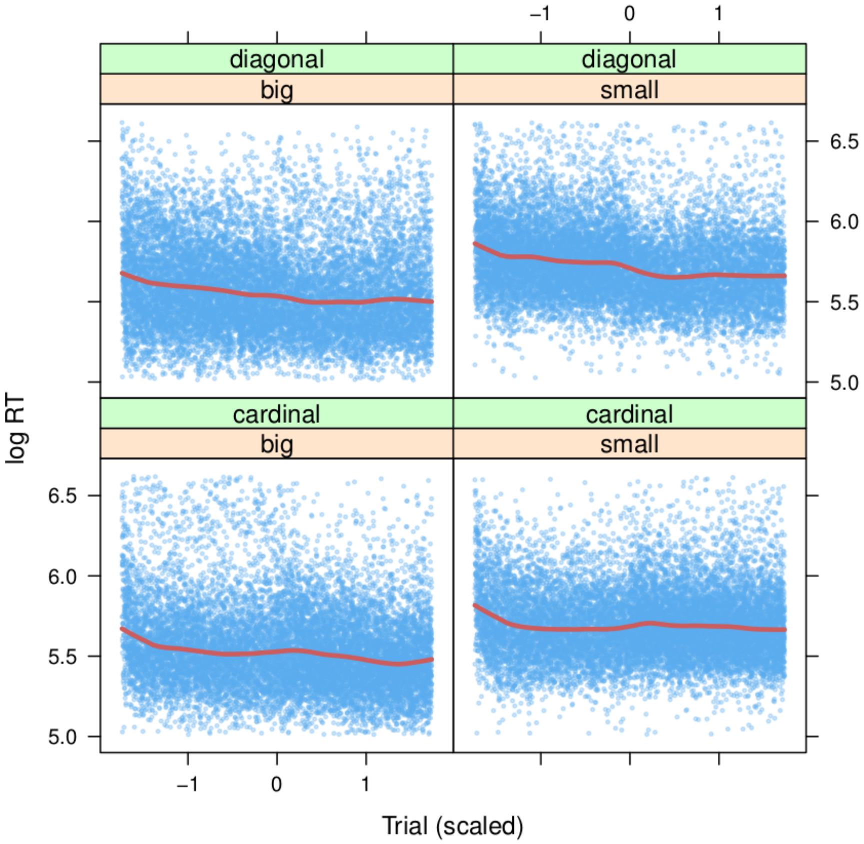

Figure 6: Loess smooths (span = 0.2) for the three-way interaction of `Trial` by `Orientation` by `Size`.

in the lower set of panels of Figure 5. These panels present scatterplots of the data points for the four subjects in the corresponding top panels, to which two smoothers have been added, a loess smoother (blue) and a smoother obtained with a generalized additive model (red). For subject 3, whose time series of responses hardly shows any autocorrelation, we observe smooths that are close to horizontal lines. As the experiment proceeds, there are only very small changes in average response time. When we move further to the right in the array of panels, temporal patterns begin to emerge. As the experiment proceeded, subjects responded more quickly. Furthermore, it appears that there may be undulations in response speed. These oscillating changes in amplitude, if real, may reflect slow changes in subjects' attention or concentration over the course of the experiment. By contrast, the general downward trend present for subjects 43, 136, and especially 123 may point to familiarization with and gradual optimization of response behavior for the task.

The presence of a potential learning effect raises the question of whether learning proceeded in the same way across the different experimental conditions. Graphical exploration suggests that the rate at which subjects respond faster over time indeed varies, specifically so across the levels of `size` and `orientation`, as shown in Figure 6. In the upper panels (`diagonal` orientation), reaction times decrease over the first half of the experiment and then level off, with a somewhat greater increase for `small` size. For `cardinal` orientation, reaction times decrease more quickly early on in the experiment, level off near the middle of the experiment, and then continue their descent for the condition with size `big`.

Within the context of the linear mixed model, the observed effects of trial can be taken into account by incorporating by-subject random slopes for `Trial`, and by allowing `Trial` to interact with `Size` and `Orientation`. As shown in Table 1, these extensions of our reference model are solidly supported by model comparisons using likelihood ratio tests. A summary of the final linear mixed model can be found in Table 3 in the appendix.

	Df	AIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
reference model	25	-25087.68	12568.84	-25137.68			
add Trial L * (sze+orn)	28	-25884.64	12970.32	-25940.64	802.96	3	< 0.0001
add Trial Q * (sze+orn)	31	-26174.41	13118.20	-26236.41	295.77	3	< 0.0001
add random slopes Trial	33	-26988.91	13527.45	-27054.91	818.50	2	< 0.0001

Table 1: Model comparisons for linear mixed models fitted to the `KKL` dataset. L: linear term of a quadratic polynomial; Q: quadratic term of this polynomial; sze: Size; orn: Orientation.

Figure 7 presents the autocorrelation functions for the residuals of this final, comprehensive, linear mixed model. Comparison with the top panels of Figure 5 shows that for subjects 136 and 123, the autocorrelation in the residuals has been reduced substantially, thanks to bringing the effects of `Trial` into the model. Nevertheless, some autocorrelation remains present.

For further reduction of autocorrelations, it is necessary to relax the assumption that the subject-specific effects of `Trial`, currently modeled by means of by-subject random intercepts and random slopes, are strictly linear. The smooths presented in the bottom panels of Figure 5 suggest that undulations may ride on top of the linear trends. What we need, then, is a way of relaxing the linearity assumption for the by-subject random effects of `Trial`. The factor smooth interaction of the generalized additive mixed model provides the required nonlinear counterpart to the combination of random slopes and random intercepts. A factor smooth for `Trial` by subject sets up a separate smooth for each level of the factor `Subject`. When we add the constraint that each smooth should

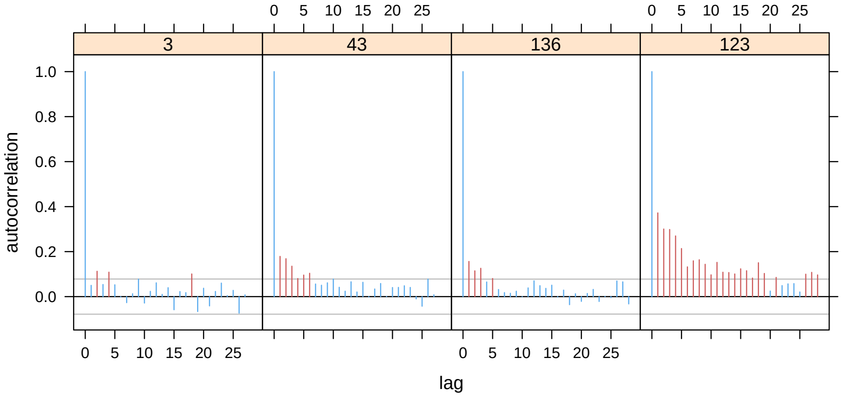

Figure 7: Autocorrelations in the residuals of the extended linear mixed model fitted to the `KKL` data.

have the same smoothing parameter, and penalize the smooths for wiggliness, thereby shrinking them towards zero, we obtain 'wiggly random effects'.

The red smooths in Figure 5 are such factor smooths. They are very similar to the loess smooths, but are slightly more sensitive to the undulations in the data. Althought it might seem there is a risk that the factor smooths are modeling noise rather than signal, this is unlikely as the factor smooths are evaluated within the general framework of the generalized additive mixed model, and hence it is possible to assess whether they contribute significantly to the model fit. By way of example, across 100 random permutations of `Trial` for the four subjects of Figure 5, a significant factor smooth was obtained in 3 instances for $\alpha = 0.05$ and for zero cases for $\alpha = 0.01$, indicative of nominal Type I error rates. This example illustrates informally that factor smooths are unlikely to find and impose nonlinear structure when there is none.

Importantly, we think the undulating random effects captured by factor smooths represent the ebb and flow of attention. They emerge not only in the present data set, but have been observed for visual lexical decision (Mulder, Dijkstra, Schreuder and Baayen, 2014) as well as for word naming and for EEG data (Baayen, van Rij, de Cat and Wood, 2016). If this interpretation is correct, penalized factor smooths are the appropriate statistical tool to use. We explicitly do not want to model these fluctuations in attention as fixed effects, because there is no reason to believe that if the experiment were replicated, a given subject would show exactly the same pattern. It is more realistic to expect that changes in attention will again be present, with roughly the same magnitude, but with ups and downs occurring at different points in time. In other words, we are dealing here with temporally structured noise, and the penalized factor smooths make it possible to bring such 'wiggly random effects' into the statistical model.

	AIC	fREML	Df	comparison with	Chisq	Df difference	Pr(> Chisq)
reference model	-26012.06	-12500.67	24				
linear model	-28047.29	-13422.47	34	reference model	921.8	10	< 0.0001
factor smooths	-30774.94	-14449.56	25	linear model	1027.1	-9	†
smooth Trial	-31040.29	-14582.62	33	factor smooths	133.1	8	< 0.0001

Table 2: Model comparison for GAMMs fitted to the KKL dataset. (The reference and linear models were refitted with GAMMs to ensure comparability across all four models.) †: the model with factor smooths has lower fREML (fast REML, see mgcv documentation) and fewer parameters, and thus is simpler and better, than the linear model.

Table 2 lists four GAMMs that we fitted to the KKL data. The first is the reference model, but refitted with GAM software (the mgcv package for R), rather than with LMM software (the lme4 package for R), with as the only change that a thin plate regression spline is used for the SOA covariate, instead of a quadratic polynomial. The second model has the same specification as the final linear mixed model (summarized in Table 3 in the appendix), but refitted with a GAMM. (These two models were refitted because both the estimation algorithms and the way in which degrees of freedom are handled differ between lme4 and mgcv.) As expected, the linear model, which includes effects for Trial, outperforms the reference model. The third model, which replaces the by-subject random intercepts and slopes by factor smooths, provides a better fit with fewer effective degrees of freedom. Addition of the three-way interaction of Trial by Size and Orientation improves the model further.

Figure 8 visualizes this three-way interaction of Trial by Orientation by Size. The effect of learning is larger for the cardinal presentation (upper panels) than for the diagonal presentation (bottom panels). For both large (left panels) and small (right panels) stimuli, we observe rapid initial learning, which levels off more for small than for large stimuli. Big stimuli with diagonal presentation elicited the most gradual accommodation pattern, with response times gradually becoming shorter.

Figure 9 clarifies that the full GAMM succeeded in further reducing the autocorrelations in the residuals. This reduction is due almost exclusively to the use of factor smooths, with only tiny amelioration by adding in the three-way interaction with Trial. This result is important for two reasons. First, it is unlikely that the removal of autocorrelation in the residuals could be accomplished by factor smooths fitting noise rather than signal. Likewise, it is unlikely that the

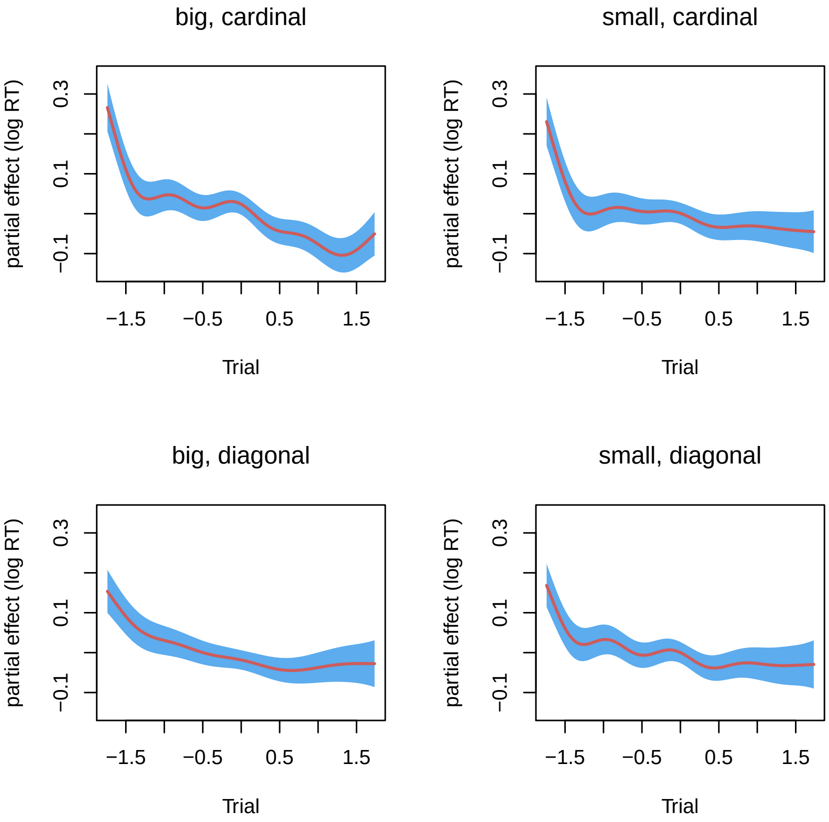

Figure 8: The `Trial` by `Orientation` by `Size` interaction in the full GAMM model for the `KKL` data.

huge reduction in AIC when going from the linear mixed model to the GAMM (2728, see Table 2) could be accomplished by just fitting noise. Second, the presence of slowly undulating processes in experimental data is a phenomenon that is itself of theoretical interest, and invites interpretation, clarification, and replication.

The remaining autocorrelations that are visible in Figure 9 are unlikely to be harmful, but to play safe, one might consider removing subject 123 from the dataset and refitting the model. Alternatively, these last remaining autocorrelations might be due to a simple AR(1) autocorrelative process in the errors, according to which the current error is equal to a proportion ρ of the preceding error plus Gaussian noise. Pinheiro and Bates (2000) and Gałecki and Burzykowski (2013) provide extensive discussion of how autocorrelation processes can be accounted for within the mixed modeling framework; Wood et al. (2015) provides technical details for GAMMs. With a mild proportionality constant $\rho = 0.15$, autocorrelations are almost completely removed. Below, we will discuss the use of this parameter in more detail. Here, we note that models from which the factor smooths

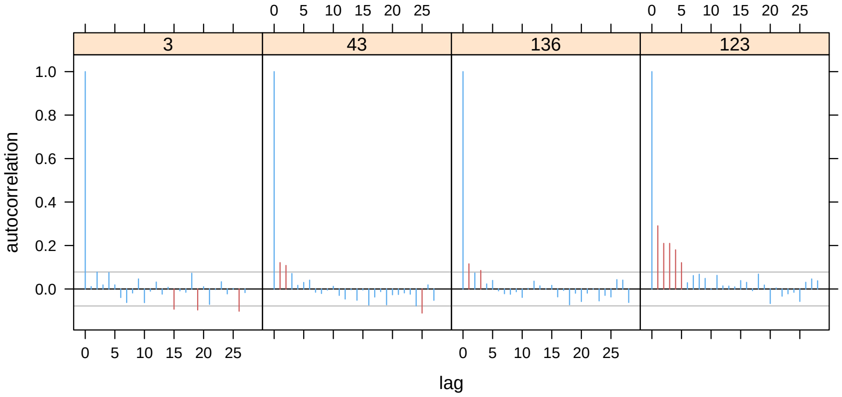

Figure 9: Autocorrelation functions for four subjects in the full GAMM fitted to the KKL data.

are removed, and for which ρ is increased, provide substantially worse fits to the data, and fail to remove substantial autocorrelational structure at longer lags. This shows that the factor smooths may be essential for bringing under statistical control a substantial part of autocorrelative structure in experimental data.

Addressing the autocorrelation issue for the KKL data set does not lead to major changes in significances and magnitudes of fixed-effect coefficients and the magnitudes of the (significant) coefficients, as reported in Bates et al. (2015) and Kliegl et al. (2015). Nevertheless, the GAMM offers enhanced insight into the data, specifically with respect to the effect of `Orientation`. In the reference model, the coefficient of this main effect was estimated at 0.041, with a standard error of 0.010 ($t = 3.9$). However, in the full linear mixed effect model, the coefficient is smaller (0.014), comes with greater uncertainty (standard error 0.09), and is not significant ($t = 1.5$). However, the final GAMM estimates the coefficient at 0.039, with a standard error of 0.016 and a t value of 2.491, reporting $p = 0.013$. Furthermore, the interaction of `Size` by `Orientation`, which is significant in the LMM, is not significant in the GAMM. Given the interaction of `Trial` by `Orientation`, the greater uncertainty about this main effect in the linear mixed model, and about the interaction with `Size`, makes perfect sense. Furthermore, as expected, the variance of the conditional modes for the by-subject random effects of `Orientation` in the reference model is larger (by a factor two) than in the final GAMM ($p < 0.0001$, F-test). Again, this makes perfect sense, as part of what originally looked like random noise linked to orientation can now be attributed to a learning effect over experimental time.

In summary, the KKL dataset is not sterile, but infected by the 'human factor'. The by-subject time series are characterized by autocorrelated errors. Unlike particles in physics, or plots of barley in agricultural experiments, human subjects are intelligent beings whose behavior is not random over time, but adaptive. The present reanalysis shows that subjects adapt in different ways to the novel manipulation of canonical versus diagonal positioning of visual stimuli, which is a theoretically fertile result. This result is not available under one-size-fits-all mechanical model selection procedures based on the a-priori assumption that the data are sterile.

4 The baldey dataset

Our second example addresses the analysis of the response latencies elicited in the auditory lexical decision megastudy of Ernestus and Cutler (2015). Ernestus and Cutler include in their statistical analysis the reaction time to the preceding trial, as a way of controlling for temporal dependencies in by-subject time series. As shown by Baayen and Milin (2010), the inclusion of preceding reaction time successfully removes a considerable amount of autocorrelation in the residuals. Ernestus and Cutler also included `Trial` as a main effect, together with by-subject random slopes for `Trial`.

In what follows, we present an analysis of the reaction times of the `baldey` data, which shows that the human factor is even stronger in these data than suggested by the analyses of Ernestus and Cutler.

In our analysis — which is far removed from a "comprehensive" analysis of this rich data set — we departed from the analyses of Ernestus and Cutler in several ways. First, we analyzed an inverse transform of the reaction times (-1000/RT) rather than a logarithmic transform, as both graphical inspection and an analysis following Box and Cox (1964) indicated the inverse transformation to better approximate normality.

Second, including the preceding reaction time is a suboptimal way of bringing autocorrelation in the errors under control. We therefore explored by-subject factor smooths in combination with the possibility of an AR(1) process in the errors.

Third, we relaxed the assumption that effects would be the same across the two genders. There are indications that males and females may be differentially sensitive to word frequency (Ullman et al., 2002; Balling and Baayen, 2008), but a gender by frequency interaction is not always found (Balling and Baayen, 2012; Tabak, Schreuder and Baayen, 2005, 2010). As the `baldey` data set combines a perfectly balanced set of subjects (10 males and 10 females) with a large number of items (2780 Dutch words), it provides an excellent testing ground for differential effects of the two genders in lexical processing.

Finally, we relaxed linearity assumptions, replacing a strictly linear mixed model by a generalized additive mixed model.

The random effects structure of the model for the reaction times for words (see Table 5 in the appendix for a statistical summary) included random intercepts for word, as well as by-word random intercepts for gender. Different words enjoy different popularity across the genders (see also Baayen, van Rij et al., 2016), and adjusting by-word intercepts for gender results in a tighter model fit. With respect to subject, we included by-subject factor smooths for session.[3] The data for this mega-study were collected over 11 sessions, and once by-subject factor smooths for session were included in the model, by-subject factor smooths for `Trial` became redundant. For subject, random slopes for the acoustic duration of the stimulus word were also well supported.

Lemma frequency (the summed frequency of a word's inflectional variants) revealed a non-linear effect that differed between females and males, as shown in the left and center panels of Figure 10. Females show a somewhat stronger frequency effect, as expected given the somewhat greater verbal skills of women compared to men (Kimura, 2000) and replicating earlier results (Ullman et al., 2002; Balling & Baayen, 2008). A novel finding is that the frequency effect appears linear for men, but shows a curvilinear pattern for women with little or no effect for very low and very high frequencies. The curvilinear pattern for females resembles that observed for the English Lexicon Project (see Figure 1), so what is remarkable here is that the men show a simpler linear effect. Possibly, both the reduced slope and the simpler functional form of the male curve is tied in with the lesser verbal skills of men.

Furthermore, the effect of the acoustic duration of the auditory stimulus showed a small but statistically well-supported modulation by `Trial`, visualized in the right panel of Figure 10. About two-thirds through an experimental session, the effect of acoustic duration decreased somewhat. This can be seen by noticing the reduced gradient for (scaled) `Trial` equal to 0.5: the number of contour lines crossed when moving horizontally across the plot, i.e., for increasing acoustic duration, is smaller than early on in the experiment.

As for the `KKL` dataset, inclusion of session and trial in the model did not absorb all autocorrelation in the residuals. With $\rho = 0.2$, the remaining autocorrelations were properly accounted

[3] Given the small number of sessions (11) and the large number of observations for each session (around 4500 for each subject), one could opt for treating session as a factor, and including by-subject random slopes for this factor. Results are very similar, with the factor smooths showing slightly more shrinkage.

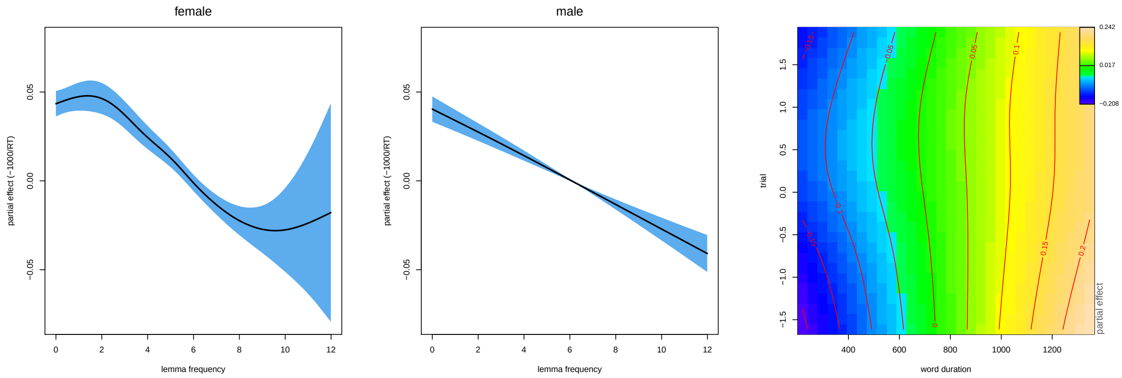

Figure 10: Interactions of `Frequency` by `Gender` (left and center panel), and of `Word Duration` by `Trial` (right panel), in the `baldey` data set.

for.

What this analysis shows is that subjects participating in an experiment with language materials bring with them their own experiences with the language, and that these specific experiences will lead to differentiated effects that for the `baldey` data set are partially differentiated by `Gender`. Furthermore, the effect of acoustic duration varied in the course of the experiment, providing further evidence for the human subject as a moving target (see also Ramscar, Hendrix, Love & Baayen, 2013; Baayen, Tomaschek, Gahl & Ramscar, 2015).

We conclude this discussion of the `baldey` dataset with observing that a stritly linear model provides an inferior fit to the data (fREML linear model: -14132.4, fREML GAMM: -15730.4, approximate test (informal because the models are not strictly nested): $\chi^2_{(5)} = 1598.0, p < 0.0001$. This linear model does not detect the interaction of frequency by gender. By imposing linearity, the nonlinear effect of frequency for females can only be accounted for by means of random intercepts and slopes, but the result is a model with a substantially worse goodness of fit. This example illustrates an important aspect of working with GAMMs: The model has to find the best balance between tracing variance back to random effects and tracing variance back to wiggly curves or (hyper)surfaces. This is why special care is required when carrying out model comparison, which we base on a comparison of fREML scores using the chi-squared test, as implemented in the `compareML` function in the `itsadug` package for `R`.

5 The poems dataset

Our final example addresses a data set previously discussed by Baayen and Milin (2010), available under the name `poems` in the `RePsychLing` package. This data set comprises a total of 275996 self-paced reading latencies from 326 subjects, for 2315 words appearing across 87 modern Dutch poems. Words are partially nested under poems. Any given subject read only a subset of poems.

Baayen and Milin included random intercepts for subject, word, and poem, as well as several by-subject and by-word random slopes for various numerical predictors. These authors sought to eliminate the problem of autocorrelated errors by including trial as a predictor, as well as the self-paced reading latency at the preceding word.

As discussed in detail by Bates et al. (2015), the model of Baayen and Milin is overparameterized with respect to its random effects structure. Given the Zipfian shape of word frequency distributions and the large numbers of words occurring only once in the corpus of poems, data are too sparse to include word as random-effect factor. Furthermore, correlation parameters for by-word random intercepts and slopes in the Baayen and Milin model were quite large, with absolute magnitudes > 0.8, often an indicator of an overparameterized model. As for the `baldey` data discussed in the preceding section, including an AR(1) process in the errors is a principled and effective solution for addressing the issue of autocorrelated errors.[4]

Within the context of the present discussion, the `poems` dataset is of interest for two reasons. First, because subjects are reading connected discourse rather than responding to unrelated isolated stimuli, the autocorrelation in their responses is much stronger than in the `KKL` and `baldey` datasets. This is illustrated in the top panels of Figure 11 for four exemplary subjects. In this dataset, there is only a handful of subjects without autocorrelations, and there are subjects with even stronger autocorrelations than the ones shown here. The second row of panels shows the corresponding scatterplots with loess and GAM smoothers. Especially for subjects 19 and 183, there are temporally concentrated spikes of long reading times that are beyond what a GAM smooth can capture. The lower set of panels illustrates the limitations of what the GAMM fitted to this data set (and described in detail in Table 7 in the appendix) can accomplish. For subject 265, the autocorrelations are properly removed, and for subject 176, the reduction in autocorrelation is perhaps satisfactory. This is not the case for subjects 19 and 183, unsurprisingly given the spiky trends in the scatterplots.

[4] Including the previous reaction time as covariate in order to reduce the autocorrelation in the error, as suggested by Baayen and Milin (2010), has many disadvantages compared to including an AR(1) process in the errors, and is not recommended.

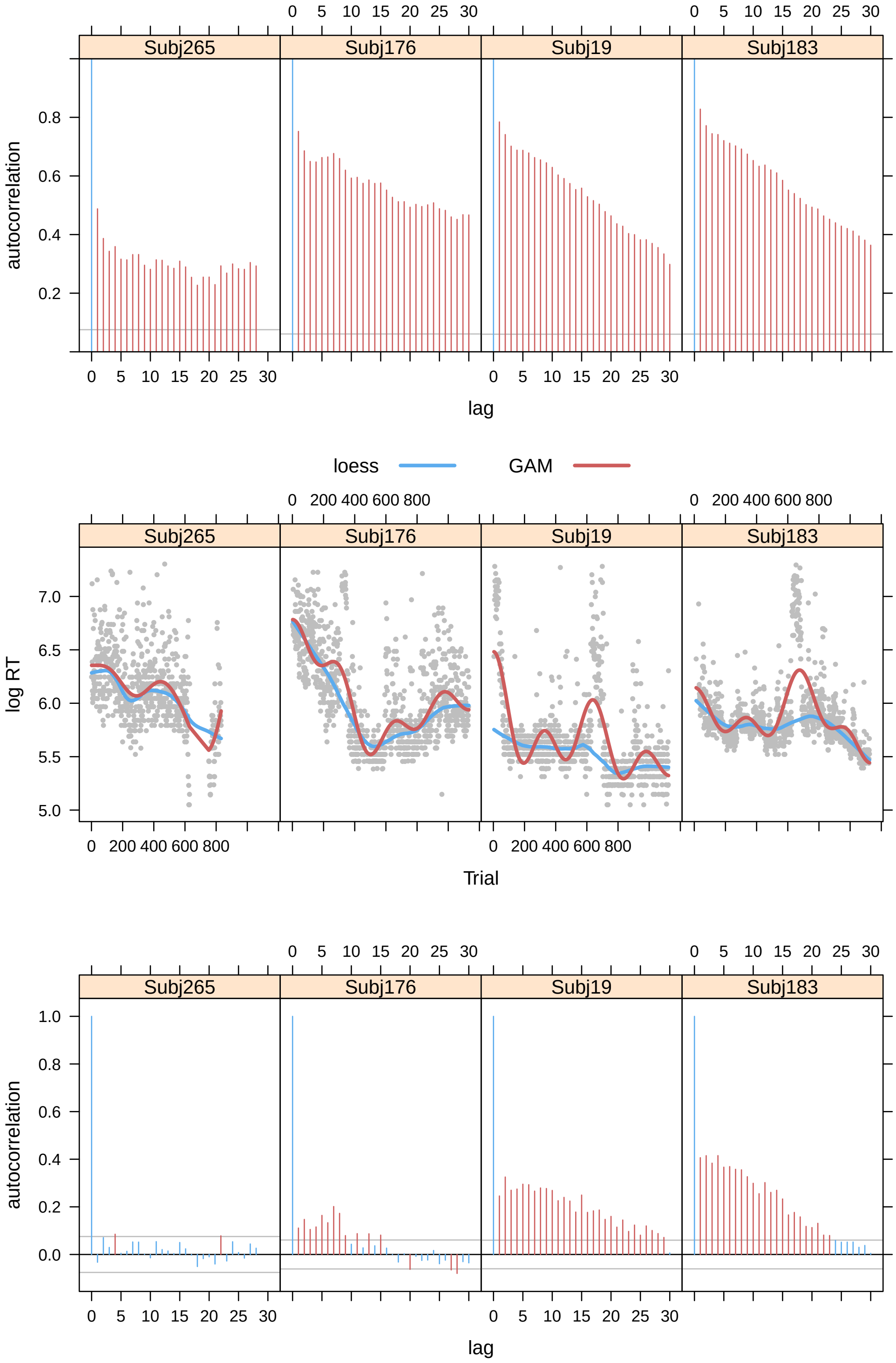

Figure 11: Autocorrelation functions for log reaction time for four exemplary subjects in the `poems` data set (upper panel), the corresponding plots (center panels) graphing log reaction time against trial, with a loess smoother (span = 0.2, in blue) and a GAM factor smooth (red), and autocorrelations in the residuals of the GAMM ($\rho = 0.3$).

Increasing ρ is not an option. As discussed in further detail in Baayen, van Rij et al. (2016), since different subjects typically emerge with different degrees of autocorrelation, one would want to adjust the ρ parameter for each individual subject. Unfortunately, it is at present not known how to achieve this mathematically within the framework of the generalized additive mixed model. As a consequence, the optimal ρ is one that strikes a balance, such that the autocorrelation for subjects with strong autocorrelation is reduced as much as possible, without introducing artificial negative autocorrelation at short lags for subjects with little or no actual autocorrelation in their residuals.

Keeping in mind the caveat that the GAMM provides an imperfect window on the complex quantitative structure of the `poems` data, it is of interest that word frequency appears to enter into a strong interaction with `Trial`. The appendix reports two models, one with a single multivariate smooth for these two predictors, and one in which their joint effect is decomposed into separate, additive, main effects of `Trial`, `Frequency`, and their interaction (see also Matuschek, Kliegl & Holschneider, 2015). These three partial effects are presented in Figure 12. We see a linear facilitatory main effect of (log-transformed) `Frequency` (left panel), a U-shaped effect of `Trial` (center panel), and an interaction that rides on top of these two main effects (right panel). The contour plot indicates that in the early trials, frequency had a more downward-sloping gradient. Later in the experiment, the effect of frequency is attenuated. The reduction in the magnitude of the frequency effect as the experiment proceeds makes sense. As subjects read through the poems selected for them, they tune in to this particular genre and its vocabulary. Words are repeated, and become more predictable as words align into sentences, and sentences into poems. As a consequence, frequency of occurrence as a contextless lexical prior becomes increasingly less informative.

We conclude with noting that all effects also receive generous support in a linear mixed effects model, but that this model lacks in precision, as indicated by model comparison: fREML linear model: 75313.32, fREML GAMM: 49307.18, approximate test (because the models are not strictly nested): $\chi^2_{(2)} = 26006.1, p < 0.0001$.

6 Regression modeling strategies

We have presented three examples demonstrating interactions of the human factor with predictors of theoretical interest. This raises the question of how to proceed with the analysis of non-sterile experimental data. In what follows, we first address this question in the context of confirmatory (or hypothesis-testing) data analysis, and then turn to exploratory (or hypothesis-generating) data

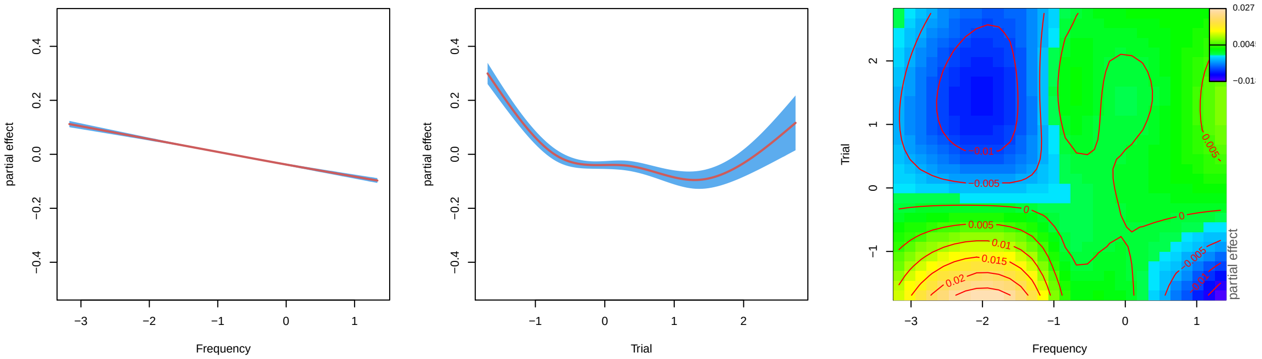

Figure 12: Partial effects for `Frequency` and `Trial` in a GAMM with an ANOVA decomposition into additive main effects and interaction, for the `poems` data set.

analysis.

6.1 Confirmatory data analysis

An excellent introduction to confirmatory multivariate data analysis is the monograph on regression modeling strategies by Harrell (2001). For clarity of exposition, we simplify analytical reality and describe the analysis as proceeding in three discrete steps. During the first step, the data are validated and explored visually, the distributions of the response variables and the distributions of the predictors are inspected, and transformed where necessary (Box & Cox, 1964). In the light of what has been learned from the initial survey of the data, including indications about non-linearities and the potential importance of covariates and possibly the presence of the human factor, a regression model can now be formulated. At the second step, the regression model is fitted to the data, and significance is assessed. This is the single and only time in the analytical process that a regression model is assessed. The third step proceeds with model criticism. At this stage, it is important to ascertain that the model fitted to the data at step 2 is indeed appropriate for the data. For a Gaussian regression model, for instance, it is important to verify that the residuals approximately follow a normal distribution, that they are independently and identically distributed, and that they do not show systematic variation with any predictors nor with the response variable. It is only when a critical term in a regression model withstands all attempts to bring it down with model criticism that one may conclude that there is reason to think that, given the simplifying assumptions that come with any regression model (see below for further discussion), a particular effect is actually supported. The size of the effect, compared to the effects of other predictors in the model, as well as the corresponding uncertainties associated with the parameter estimates, will be essential for the assessment of the scientific importance of this support. Importantly, the parameters in the model should be meaningful, at two levels. Mathematically, parameters should be properly estimable and interpretable. Furthermore, at the level of domain knowledge, all parameters should be theoretically interpretable. For instance, by-subject random intercepts in a regression model fitted to a reaction time study are interpretable as a random variable placing subjects on a scale from fast to slow responders, and by-subject factor smooths for experimental time are interpretable as reflecting the ebb and flow of attention.

Since a confirmatory analysis allows one, and only one, statistical test for the evaluation of a specific hypothesis, it is of crucial importance that this test is based on models that are not too complex to be estimable, on models that are properly interpretable, and on models that take the human factor into account if it is present. How then might one proceed under these stringent boundary conditions?

At first sight, it might be argued that a model should be fit to the data that is as complex as possible, a model that takes all possible contingencies into account that might put the critical model parameter in jeopardy. Thus, one might think that it is straightforward to enrich a maximal linear mixed model with predictors targeting the human factor. Unfortunately, once one enters the nonlinear world, this is even less advisable than for the linear world, for a variety of reasons.

First, more elaborate models can quickly become very difficult to understand. By way of example, a model with a four-way interaction of `Word Duration`, `Session`, `Frequency` and `Trial` improves substantially on the model presented above for the `baldey` data set ($\chi^2_9 = 24.35, p < 0.0001$). But what we learn from this four-way tensor product is unclear. In 1959, Sigmund Koch wrote that "Psychology [is] unique in the extent to which its institutionalization preceded its content and its methods preceded its problems." (Koch, 1959, p. 783). Whereas this may not be true for all areas of psychological science (e.g., some areas of vision research), it certainly applies to the domain of lexical processing. Here, GAMMs will often be informative about possible structure in experimental data that is far beyond what current theories can explain or predict. For the `baldey` data, we deliberately avoid a 'maximal' model, as, given current knowledge, it is unclear whether such a model would contribute to understanding the data.

Second, when we make use of a factor smooth with shrinkage to fit nonlinear by-subject trends over experimental time, we are making many simplifying assumptions, among which (i) that all subject smooths are can be captured with the same smoothing parameter, and (ii) that these temporal trends do not interact with other predictors in the model, whether factorial (say, a priming condition) or metric (say, frequency or valence). These assumptions may or may not be valid, but it typically does not make much sense to aim for a complex model term such as a tensor product smooth for trial by frequency by valence by priming by subject, with our without shrinkage. Again, given our current state of knowledge, such complex models, if at all estimable, will typically be very difficult to interpret.

Third, fitting a complex generalized additive mixed model is not a trivial issue, and for results to be sensible, it is crucial to avoid random effects structure that is internally collinear (see Bates et al., 2015, for detailed discussion). In general, as observed by Wood (documentation for `gam.selection` in the `mgcv` package),

> The more thought is given to appropriate model structure up front, the more successful model selection is likely to be. Simply starting with a hugely flexible model with 'everything in' and hoping that automatic selection will find the right structure is not often successful.

Would researchers come to 'wrong' conclusions if they analyze data simply using maximal linear models, without taking the human factor into account, without paying attention to whether the model is overfitting the data with mathematically uninterpretable parameters (Lele, Nadeem & Schmuland, 2012; Bates et al., 2015), and accepting less than nominal power (Matuschek, Kliegl, Vasishth, Baayen & Bates, 2015)? The problem here is that low-hanging fruit is easily plucked, often by simple linear models without any random effects. The devil is in the details. Significance of factorial contrasts may not change, or may not change by much, when the human factor is taken into account. For the `KKL` data set, we showed that a full mixed model would lead to the conclusion that the main effect of `Orientation` is not significant, whereas a model that takes the human factor into account suggests otherwise. Furthermore, the maximal model suggests that the interaction of `Size` and `Orientation` is significant, but the GAMM with predictors for the human factor indicates that there is no support whatsoever for such an interaction. Details change, but the three-way interaction of `Orientation` by `Size` by `Gravitation` remains. Do the details matter?

If details don't matter, in many cases the analyst will be well off with a simple linear model, even a linear model without random effects. Often, conclusions about 'significance' do not change when data sets are analysed with much simpler models. However, when a simple linear model produces the same verdict on significance as a linear mixed model, or a linear mixed model provides the same verdict of significance as a generalized additive mixed model, this does not mean that the more complex modeling technique is not required. Even when conclusions about significance of predictors do not change for observed examples, they might change substantially for as yet unobserved examples. More importantly, random slopes and random intercepts typically modulate effect sizes and degrees of uncertainty. Especially when it comes to prediction, more precise estimates that take into account subject and item variability are invaluable. Similarly, taking into account nonlinearities and human factors may modulate conclusions about effect sizes and the precise nature of functional relations. For the `KKL` dataset, we observed a significant partial effect of `Orientation`, modulated by interactions with far smaller effect sizes. The partial effect of `Orientation` is of theoretical significance, and it therefore is important to use a modeling strategy that makes its partial effect visible.[5]

Given that a maximal model approach provides the false security of a comfort blanket, the question remains how one might proceed under the stringent boundary conditions of confirmatory data analysis. All we can do to answer this question is present examples of how one might proceed. Consider a chronometric experiment with subjects and items. Inspired by Harrell (2001), one possible way to proceed could be as follows. As a first step, following data validation, exploratory visualization is carried out. At this stage, non-parametric scatterplot smoothers could be used to probe for the presence of by-subject trends over experimental time. Furthermore, the autocorrelation function could be obtained for the response variable, in order to assess what value of ρ might be required. However, because the temporal autocorrelation can be due to the combined presence of an AR(1) process in the errors and subject-specific trends in experimental time, the autocorrelation

[5] It might be argued we have not shown that addressing autocorrelated errors changes conclusions about the effect of `Orientation`. The argument runs as follows. The main effect of a predictor X that interacts with a predictor Y specifies the effect of X when $Y = 0$. Since `Orientation` interacts with `Trial`, the main effect of `Orientation` specifies its effect when `Trial` $= 0$ (i.e., in the middle of the experiment, since `Trial` was scaled). From this, it would follow that we have no case to argue that it is the explicit treatment of autocorrelated errors that has changed the apparent conclusions from the model about the effect of `Orientation`. This argument misses three important points.

First, autocorrelation in the errors is addressed in part by the by-subject factor smooths for `Trial`, and in part by the autocorrelation parameter ρ. It is the combination of the two that leads to different conclusions about the effect of `Orientation`. Second, as explained in section 2, main effects in a model with interactions are crucial for properly calibrating the wiggly curves for individual factor levels with respect to the intercept. They are an essential part of the model. Changing orientation from cardinal to diagonal implies a modulation of the intercept by 0.078 units on the -1000/RT scale. This change effects all trials for the relevant factor level. A maximal LMM estimates the effect to be much smaller (0.028) and not significant. In other words, the maximal LMM underestimates the distance between the relevant curves. Third, as a consequence, the maximal LMM provides a warped perspective of the magnitute of the effect of `Orientation` vis-à-vis the other predictors in the model.

function for the response may overestimate the value of ρ when by-subject trends in experimental time are in fact present. Therefore, it may be preferable to fit an intercept-only GAMM with factor smooths for subject and random intercepts for item to the data, with as only aim to detect with more precision the extent to which the human factor is present, to determine whether it is necessary to include by-subject factor smooths, and to obtain an estimate for the autocorrelation parameter ρ. If there is no clear evidence for the human factor, a linear mixed model is an excellent choice, otherwise, a GAMM is preferable, with an autocorrelation parameter for the AR(1) process in the errors set close to the autocorrelation at lag 1 observed for the intercept-only model.

Then, at step two of the analysis, a model could be fitted with all relevant fixed-effect parameters added in, but without any further random effects and random slopes. Significance of the critical predictor can now be assessed through model comparison with a simpler model from which the critical predictor, or the relevant critical interaction with this predictor, is removed. Importantly, this is the single and only time at which significance is assessed.

The final step proceeds to model criticism. At this step, the question is whether significance (if established) will survive removal of overly influential outliers, addition of further random-effect parameters (specifically, and importantly, random effects or random slopes for the critical model terms), adjustment of the autocorrelation parameter if necessary, and inclusion of interactions with human-factor variables. Bootstrap validation is also worth considering at this step. If an effect withstands model criticism, it can be reported as significant with the p-value obtained at step 2, otherwise, it should be reported as not significant. This confirmatory modeling strategy has the advantage that models that overfit the data with meaningless parameters are avoided. As pointed out by Lele et al. (2012), "Whenever mixed models are used, estimability of the parameters should be checked before drawing scientific inferences or making management decisions". Since meaningless parameters can arise even under convergence (see Bates et al., 2015), the analyst may want to minimize the risk of running into this situation specifically when significance is assessed in a confirmatory context.

Importantly, there are other strategies that could be followed, such as starting with a model including all random intercepts and all random effects and slopes, while leaving out correlation parameters (Bates et al., 2015). Here, a confirmatory setting takes the significance of the pertinent predictor as the outcome of interest, and subsequent model criticism is carried out to ensure that this significance is trustworthy. If it turns out that the model is too complex to be supported by the data, the analyst may want to refit a simpler and better justified model, in which case the analysis has become exploratory — it is only as long as significance is evaluated once, and once only, with subsequent model criticism to ensure support for the original significance test, that the analysis is a proper confirmatory analysis. What specific strategy is followed, and we have given only two of many possible strategies, is, to a large extent, a matter of taste, so it would make sense to report the details of the strategy that was actually followed for a given confirmatory analysis. In any case, the most transparent way to proceed is to release all data and code with the published paper, so that readers have the option to draw their own conclusions from the data.

6.2 Exploratory data analysis

One of the causes of the deplorable rate of replication failure for psychology (and many other fields of inquiry) is that confirmatory data analysis is seen as vastly superior to exploratory data analysis, and that as a consequence, the results of many exploratory data analyses are presented as if they were the result of confirmatory analysis.

All models reported in the present study are the result of data exploration with step-by-step theory-driven incremental model building (for examples of such an approach, see Pinheiro & Bates, 2000; González, De Boeck, Tuerlinckx et al., 2014). The t and p-values reported in the appendix, therefore, are indicators of surprise and should not be taken at face value as exact probabilities. (Note, however, that the crucial probabilities for the human factor are so small, often $<2e$-16, that they may be expected to survive substantial correction for multiplicity.) We believe that exploratory data analysis is of great importance for those domains of inquiry where explicit and mathematically precise theories are lacking. In these areas of inquiry, results of experiments can be completely opposite to what was expected, even though in hindsight, they may make perfect sense (see, e.g., the anti-frequency effect in Figure 3, for a computational model based on discriminative learning that captures this effect, see Pham & Baayen, 2015).[6] Importantly, in an exploratory setting, one can actually learn from the data, in a multivariate setting often in many dimensions simultaneously, instead of receiving only confirmation (or disconfirmation) of a single hypothesis.

Also in an exploratory setting, model criticism is absolutely essential. An effect is worth taking seriously only if it withstands truly serious attempts to bring it down. Researchers may want to complement exploratory regression analysis with techniques from machine learning such as random

[6] The review process forced a presentational style on this study in which the experimental results are reported as being predicted by discriminative learning theory, whereas the unexpected experimental results preceded in time the subsequent modeling with discriminative learning.

forests (Breiman, 2001; Strobl, Malley & Tutz, 2009) or gradient boosting machines (Friedman, 2001; Chen, He & Benesty, 2015) to obtain an independent assessment of variable importance that is orthogonal to the exploration of the data with regression modeling. These techniques are not plagued by issues of collinearity (Wurm & Fisicaro, 2014), and they do not require any kind of model selection (for an example, see, e.g., Baayen, Milin & Ramscar, 2016, for application of random forests for this purpose).

6.3 A gold standard?

The Open Science Collaboration (2015) reported the results of an extensive series of replication studies, and documented a 50% drop in observed effect sizes and a drop from 97% to 36% of significant results. Publication bias (Francis, 2012), non-random selection of stimuli and subjects (Forster, 2000; Sander and Sanders, 2006; Francis, 2013), societal changes (Ramscar, Shaoul and Baayen, 2015) and especially lack of power (Button et al., 2013, see also Westfall, Kenny and Judd, 2014) are major concerns. Clearly, many of the results in the published literature are but moving shadows of the true effects, which raises the question how to escape from the current methodological cave.

Barr, Levy, Scheepers and Tily (2013) proposed a new standard for the statistical analysis of experimental data. Based on a series of simulation studies, they argued that anti-conservative p-values are best avoided by fitting linear mixed effects models with all variance components included that could in principle be non-zero given the experimental design. The simulations on which Barr et al. (2013) base their recommendation make very specific assumptions:

> We assumed no list effect; that is, the particular configuration of 'items' within a list did not have any unique effect over and above the item effects for that list. We also assumed no order effects, nor any effects of practice or fatigue. (p. 264)

Given our limited understanding of the human factor, this simplification is understandable. However, this very simplification invalidates their simulation design as a foundation for a general gold standard. We have shown that the human factor may interact with key predictors, and that it may lead to different conclusions about the details of their effects. We note here that it is not possible to prove anything via simulation. One can show trends, but extrapolation is not recommendable, and should never be the foundation for a gold standard for a field of inquiry. Given that our understanding of the human factor in psycholinguistic experiments is poor, it is unlikely that simulation studies including human factors are feasible or worthwhile as a way forward.

Since the proposed standard of Barr et al. is overly conservative with an unnecessary loss of power (Matuschek, Kliegl, Vasishth et al., 2015), and since it comes with the risk of basing conclusions on mathematically ill-defined models that overfit the data (Lele et al., 2012; Bates et al., 2015), this proposed gold standard has the unfortunate side-effect of locking analytical practice in a methodological cage of shadows from which the true structure of experimental data, rich and fertile in perhaps unexpected ways, cannot be fully appreciated.

Recommendations such as the one of Barr et al., which once in the literature quickly rise to the status of rules enforced in the review process with an iron fist, have the unfortunate side effect of diverting attention from the model, the balance of forces within the model, the uncertainties associated with the model, and its inevitable weaknesses, to the so fervently desired but so over-valued p-value. A one-size-fits-all rule for obtaining such p-values might seem attractive as the only way to obtain an 'objective' procedure evaluating experimental effects. However, irrespective of whether analysis is exploratory or confirmatory, in the modern age, objectivity can be achieved in a much more direct way. Whereas before the advent of the internet, reporting p-values in printed journals was the only way to make an argument that a particular effect is present, current communication technology makes it possible to publish not only p-values but the data themselves, using platforms such as the Mind Research Repository at http://openscience.uni-leipzig.de/index.php/mr2 or the Open Science Framework (https://osf.io/), where data can be made available together with details on the statistical analysis. With the data out in the open, readers will not only be able to evaluate for themselves the appropriateness of the analyses reported in a journal, but opportunities are created for improved analyses, either with current or with future statistical software. In this way, a degree of objectivity can be reached that goes far beyond what can be obtained by mechanical procedures and the considerable risk of associated artifacts. Furthermore, through meta-analysis, one can build on previous findings.

7 Discussion

Even though there is considerable awareness in the literature on memory and language that time series of behavioral responses are not independent (Broadbent, 1971; Welford, 1980; Sanders, 1998; Taylor & Lupker, 2001; De Vaan et al., 2007; Baayen & Milin, 2010; Masson & Kliegl, 2013), the fact that this interdependence has far-reaching consequences for the statistical analysis of experimental data has not received systematic reflection. We have reported three data sets in which inter-trial dependencies are present. Unlike molecules or barley, the units from which response variables in psychology are harvested are intelligent beings that constantly adapt to their environment. Humans learn. They get tired. Their attention wanes, and then is refocused on the task. When attentional and adaptative processes are demonstrably present in experimental data, and interact with predictors of theoretical interest, it is advisable to bring these processes under statistical control. Failure to take the human factor into consideration comes with the risk of misunderstanding the finer details of the quantitative structure of the data and the extent to which this structure is shaped by the predictors of interest.

We have introduced the generalized additive mixed model as an extension of the linear mixed model that makes it possible to bring the human factor into the statistical model, and to take the statistical sting out of the autocorrelations in the residual error. We do not wish to claim that with GAMMs researchers will finally emerge from the cave of shadows and apprehend the true effects themselves. But we are finding GAMMs helpful in sharpening the contours and outlines of these effects. For almost all data sets that we have investigated with GAMMs, we have obtained better fits by including by-subject factor smooths for trial. We also demonstrated interactions of trial with predictors of interest, a first step towards a better understanding of the human factor in lexical processing.

As any statistical model, GAMMs build on assumptions that are hoped to be reasonable, but of which we often know they involve substantial simplifications. When using GAMMs, it is important to have good coverage of the covariate space, especially when using generalized models with Poisson or binomial families, and highly irregular regression surfaces with highly localized effects should be treated with caution.[7] Unlike the linear mixed model, the GAMM as implemented in the `mgcv` package does not offer the possibility to test for correlation parameters in the random effects. Analysts used to the speed with which the software of the fourth author fits LMMs will find the speed with which GAMMs with complex random effects structure converge excruciatingly slow, the price paid for not imposing prior constraints on the random effects structure. Furthermore, the

[7] GAMs do not require especially large data sets. Examples from Wood (2006) include a data set with only 31 observations on girth, height, and volume for black cherry trees, where a thin plate regression spline for girth appears justified, and a data set with 634 observations on mackerel eggs where a thin plate regression spline for longitude and latitude is part of the model specification. Of course, when there is only a handful of distinct values for a given covariate, a smooth for that covariate will not make sense. The smoothers in the `mgcv` package typically produce an error message for such cases.

penalized factor smooths that we have used assume a common smoothing parameter for all subjects, which may be true, but may also be incorrect. The factor smooths build on spline theory, but when time series of reaction times are spiky instead of smoothly wiggly, splines are better than nothing, but certainly not perfect. To this list of limitations we can also add that the adjusting of the errors for an AR(1) autocorrelative process will often be too simplistic in two ways. First, as we have demonstrated, ρ should ideally vary with participant, which is currently not possible with the `gam` (and `bam`) functions of the MGCV package (but progress may be possible here by going fully Bayesian, see Wood, 2016). Second, there is no guarantee that the autocorrelative process is a simple AR(1) process — dependencies may well stretch further back in time. Nevertheless, we believe that with GAMMs, researchers have a tool in hand with which the real-life complexities with which psycholinguistic data may be infected can start to be investigated.

We conclude with some reflections on parsimony in regression modeling. George Box is well known for stating that all models are wrong, but that some are useful (Box, 1979). With respect to model parsimony, he noticed that

> Since all models are wrong the scientist cannot obtain a "correct" one by excessive elaboration. On the contrary following William of Occam he should seek an economical description of natural phenomena. Just as the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparameterization is often the mark of mediocrity. (Box, 1976, p. 792)

When working with GAMMs, models of considerable complexity can be fit to experimental data sets. Since language is a complex cognitive skill serving users interacting in highly complex and technologically advanced societies, it is perhaps not surprising that statistical models may need to be more complex than previously thought when it comes to taking into account the human factor. Yet, it is important to resist the temptation to go 'maximal' (as advised by Barr et al., 2013, in the context of the linear mixed model). Within the boundary conditions of not overparameterizing, of properly balancing Type I and Type II error rates, and of having a model with meaningful and interpretable parameters, the challenge is to find the right balance between model simplicity and faithfulness to the data. Within the context of confirmatory data analysis, a possible modeling strategy might be to keep the model lean, with factor smooths and an autocorrelation parameter included when there appears to be evidence for the human factor, and to shift the burden of securing a reasonable degree of confidence about a critical effect to model criticism. It must be acknowledged that an important achievement of the Barr et al. (2013) paper was to raise awareness about the dangers of fitting overly simple linear mixed models. However, their recommendation (summarized in the title of their paper) to "keep it maximal", crosses over to such an extreme position as to be untenable in most realistic statistical modeling settings.

Within the context of exploratory data analysis, incremental hypothesis-driven model building can yield an accumulation of valuable insights. Importantly, theory and experience should guide model building, counterbalancing faithfulness to the data with a drive for simplicity. For the `KKL` data set, for instance, the three-way interaction of `Orientation` by `Size` by `Trial` was explored because we thought it was conceivable that within-experiment learning and adaptation might vary with the difficulty of the different testing conditions. It is possible that the details of this interaction vary from subject to subject. Without hypotheses about what might drive individual differences in a four-way interaction with subject, such an interaction, even if it were estimable, would be extremely difficult to understand. For the `baldey` data set, we considered the possibility that listeners might adapt to the speech rate of the speaker as the experiment proceeded. However, we refrained from discussing a model in which `Trial` and `Word Duration` entered into further interactions with `Session` and `Frequency`, even though such a four-way interaction, which is estimable, improves substantially on the model presented above for the `baldey` data. This more complex model might be capturing something real, but without guidance from theory, and without further support from replication studies, it is unclear what the benefits of such a complex model might be. For the `poetry` data set, the model reported in Baayen and Milin (2010) turned out to be overparameterized, the data for words being too sparse to allow inclusion of by-word random intercepts and slopes. Therefore, our GAMM for the `poetry` data did not include random effects for words. In short, also in the context of exploratory data analysis, parsimony is a virtue, not a vice.

The probability of objectivity and replicability in data analysis, whether exploratory or confirmatory, is likely to be enhanced when researchers make their data and analyses available to the research community. Just by itself, making the data available is a substantial deterrent for reporting results of fishing expeditions: Anyone with access to the data will immediately spot that the model published is an implausible one. Publication of data will also render fishing expeditions across multiple methods for analyzing the data less attractive. For instance, reporting an F1+F2 analysis (Forster & Dickinson, 1976) that supports significance after first having observed a lack of significance in a LMM is straightforward to detect. Rather than seeking to guarantee objectivity through rigid methodological one-size-fits-all prescriptions, we think open science, combined with responsible statistical analysis, is the way forward out of the cave of shadows.

References

Arnon, I. & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. *Cognition*, *122*(3), 292–305. doi:10.1016/j.cognition.2011.10.009

Baayen, R. H. (2010). *languageR: data sets and functions with "analyzing linguistic data: a practical introduction to statistics"*. R package version 1.4.1. Retrieved from http://CRAN.R-project.org/package=languageR

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

Baayen, R. H., Milin, P. & Ramscar, M. (2016). Frequency in lexical processing. *Aphasiology*, 1–47.

Baayen, R. H., Tomaschek, F., Gahl, S. & Ramscar, M. (2015). The Ecclesiastes principle in language change. In M. Hundt, S. Mollin & S. Pfenninger (Eds.), *The changing English language: Psycholinguistic perspectives* (to appear). Cambridge, UK: Cambridge University Press.

Baayen, R. H., van Rij, J., de Cat, C. & Wood, S. N. (2016). Autocorrelated errors in experimental data in the language sciences: Some solutions offered by generalized additive mixed models. In D. Speelman, K. Heylen & D. Geeraerts (Eds.), *Mixed effects regression models in linguistics* (to appear). Berlin: Springer.

Balling, L. & Baayen, R. H. (2008). Morphological effects in auditory word recognition: Evidence from Danish. *Language and Cognitive Processes*, *23*, 1159–1190.

Balling, L. & Baayen, R. H. (2012). Probability and surprisal in auditory comprehension of morphologically complex words. *Cognition*, *125*, 80–106.

Balota, D., Cortese, M., Sergent-Marshall, S., Spieler, D. & Yap, M. (2004). Visual word recognition for single-syllable words. *Journal of Experimental Psychology: General*, *133*, 283–316.

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), 255–278.

Bates, D., Kliegl, R., Vasishth, S. & Baayen, R. H. (2015). *Parsimonious mixed models*. Retrieved from http://arxiv.org/abs/1506.04967

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*, 791–799. doi:doi:10.1080/01621459.1976.10480949

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in statistics* (pp. 201–236). Academic Press.

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

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Broadbent, D. (1971). *Decision and stress.* New York: Accademic Press.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376.

Chen, T., He, T. & Benesty, M. (2015). *Xgboost: extreme gradient boosting.* (R package version 0.4-0). Retrieved from https://github.com/dmlc/xgboost

De Vaan, L., Schreuder, R. & Baayen, R. H. (2007). Regular morphologically complex neologisms leave detectable traces in the mental lexicon. *The Mental Lexicon*, *2*, 1–23.

Dery, J. E. & Pearson, H. (2015). On experience-driven semantic judgments: A case study on the oneiric reference constraint. In *Proceedings of Quantitative Investigations in Theoretical Linguistics, Tübingen, 2015.*

Egly, R., Driver, J. & Rafal, R. D. (1994). Shifting visual attention between objects and locations: Evidence from normal and parietal lesion subjects. *Journal of Experimental Psychology: General*, *123*, 161–177.

Ernestus, M. & Cutler, A. (2015). BALDEY: A database of auditory lexical decisions. *The Quarterly Journal of Experimental Psychology*, (ahead-of-print), 1–20.

Forster, K. (2000). The potential for experimenter bias effects in word recognition experiments. *Memory & Cognition*, *28*, 1109–1115.

Forster, K. & Dickinson, R. (1976). More on the language-as-fixed effect: Monte-Carlo estimates of error rates for F_1, F_2, F', and *min*F'. *Journal of Verbal Learning and Verbal Behavior*, *15*, 135–142.

Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review*, *19*(6), 975–991.

Francis, G. (2013). Publication bias in “Red, Rank, and Romance in Women Viewing Men” by Elliot et al. (2010). *Journal of Experimental Psychology: General*, *142*(1), 292–296.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232.

Gałecki, A. & Burzykowski, T. (2013). *Linear mixed-effects models using R: A step-by-step approach.* Springer Science & Business Media.

González, B., De Boeck, P., Tuerlinckx, F. et al. (2014). Linear mixed modelling for data from a double mixed factorial design with covariates: a case-study on semantic categorization response times. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *63*(2), 289–302.

Harrell, F. (2001). *Regression modeling strategies*. Berlin: Springer.

Hastie, T. & Tibshirani, R. (1990). *Generalized Additive Models*. London: Chapman & Hall.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer.

Kimura, D. (2000). *Sex and Cognition*. Cambridge, MA: The MIT Press.

Kleinschmidt, D. F. & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148.

Kliegl, R., Kuschela, J. & Laubrock, J. (2015). Object orientation and target size modulate the speed of visual attention. *Manuscript, University of Potsdam.*

Kliegl, R., Wei, P., Dambacher, M., Yan, M. & Zhou, X. (2011). Experimental effects and individual differences in Linear Mixed Models: Estimating the relationship between spatial, object, and attraction effects in visual attention. *Frontiers in Psychology*, *1*, 1–12.

Koch, S. E. (1959). *Psychology: A study of a science.* McGraw-Hill.

Lele, S. R., Nadeem, K. & Schmuland, B. (2012). Estimability and likelihood inference for generalized linear mixed models using data cloning. *Journal of the American Statistical Association.*

Lin, X. & Zhang, D. (1999). Inference in generalized additive mixed models using smoothing splines. *JRSSB*, *61*, 381–400.

Marra, G. & Wood, S. N. (2012). Coverage properties of confidence intervals for generalized additive model components. *Scandinavian Journal of Statistics*, *39*, 53–74.

Marsolek, C. J. (2008). What antipriming reveals about priming. *Trends in Cognitive Science*, *12*(5), 176–181. doi:10.1016/j.tics.2008.02.005

Masson, M. E. & 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.

Matuschek, H., Kliegl, R. & Holschneider, M. (2015). Smoothing spline ANOVA decomposition of arbitrary splines: An application to eye movements in reading. *PLoS ONE*, *10*(3).

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, R. H. & Bates, D. (2015). Balancing Type I Error and Power in Linear Mixed Models. Retrieved from http://arxiv.org/abs/1511.01864

Mulder, K., Dijkstra, T., Schreuder, R. & Baayen, R. H. (2014). Effects of primary and secondary morphological family size in monolingual and bilingual word processing. *Journal of Memory and Language*, *72*, 59–84.

Nychka, D. (1988). Bayesian confidence intervals for smoothing splines. *Journal of the American Statistical Association*, *83*, 1134–1143.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Pham, H. & Baayen, R. H. (2015). Vietnamese compounds show an anti-frequency effect in visual lexical decision. *Language, Cognition, and Neuroscience*. doi:DOI:10.1080/23273798.2015.1054844

Pinheiro, J. C. & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. Statistics and Computing. New York: Springer.

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, *32*, 3–25.

R Development Core Team. (2015). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from http://www.R-project.org/

Ramscar, M., Hendrix, P., Love, B. & Baayen, R. H. (2013). Learning is not decline: The mental lexicon as a window into cognition across the lifespan. *The Mental Lexicon*, *8*, 450–481. doi:10.1075/ml.8.3.08ram

Ramscar, M., Shaoul, C. & Baayen, R. H. (2015). *Why many priming results don't (and won't) replicate: a quantitative analysis*. Manuscript, University of Tübingen. Retrieved from http://psych.stanford.edu/~michael/papers/Ramscar-Shaoul-Baayen_replication.pdf

Ramscar, M., Yarlett, D., Dye, M., Denny, K. & Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science*, *34*(6), 909–957. doi:10.1111/j.1551-6709.2009.01092.x

Sander, P. & Sanders, L. (2006). Rogue males: Sex differences in psychology students. *Electronic Journal of Research in Educational Psychology*, *8*(4), 85–108.

Sanders, A. (1998). *Elements of human performance: Reaction processes and attention in human skill*. Mahwah, New Jersey: Lawrence Erlbaum.

Strobl, C., Malley, J. & Tutz, G. (2009). An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests. *Psychological methods*, *14*(4), 26.

Tabak, W., Schreuder, R. & Baayen, R. H. (2005). Lexical statistics and lexical processing: Semantic density, information complexity, sex, and irregularity in Dutch. In S. Kepser & M. Reis (Eds.), *Linguistic Evidence — Empirical, Theoretical, and Computational Perspectives* (pp. 529–555). Berlin: Mouton de Gruyter.

Tabak, W., Schreuder, R. & Baayen, R. H. (2010). Producing inflected verbs: A picture naming study. *The Mental Lexicon*, *5*(1), 22–46.

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

Ullman, M. T., Estabrooke, I. V., Steinhauer, K., Brovetto, C., Pancheva, R., Ozawa, K. & Mordecai, K. and Maki, P. (2002). Sex differences in the neurocognition of language. *Brain and Language*, *83*, 141–143.

van Rij, J., Baayen, R. H., Wieling, M. & van Rijn, H. (2015). itsadug: Interpreting time series, autocorrelated data using GAMMs. R package version 1.0.3.

Welford, A. (1980). Choice reaction time: Basic concepts. In A. Welford (Ed.), *Reaction times* (pp. 73–128). New York: Accademic Press.

Westfall, J., Kenny, D. A. & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*(5), 2020.

Wood, S. N. (2006). *Generalized Additive Models*. New York: Chapman & 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*, 3–36.

Wood, S. N. (2013). On p-values for smooth components of an extended generalized additive model. *Biometrika*, *100*, 221–228.

Wood, S. N. (2016). Just Another Gibbs Additive Modeller: Interfacing JAGS and mgcv. *arXiv preprint arXiv:1602.02539v1*.

Wood, S. N., Goude, Y. & Shaw, S. (2015). Generalized additive models for large data sets. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *64*(1), 139–155.

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

Zhou, X., Chu, H., Li, X. & Zhan, Y. (2006). Center of mass attracts attention. *Neuroreport*, *17*, 85–88.

Acknowledgements

This research was supported by an Alexander von Humboldt Professorship awarded to the first author. We are indebted to Simon Wood for invaluable discussion and for the many extensions he added to the `mgcv` package to facilitate the analysis of (psycho)linguistic data.

Appendix

The KKL dataset

random effects

Groups	Name	Std.Dev.	Corr
subj	Intercept	0.161766	
	Trial	0.051180	-0.273
subj.1	spt	0.066809	
subj.2	grv	0.033814	
subj.3	obj	0.025609	
subj.4	orn	0.076247	
subj.5	spt_orn	0.033226	
Residual		0.185591	

fixed effects

	Estimate	Std. Error	*t* value
Intercept	5.659	0.018	322.91
sze	0.184	0.035	5.27
spt	0.074	0.008	9.62
obj	0.043	0.005	9.41
grv	-0.001	0.005	-0.17
orn	0.014	0.009	1.51
Soa L	-0.010	0.001	-12.56
Soa Q	0.019	0.001	20.61
sze:spt	0.048	0.015	3.14
sze:obj	-0.012	0.009	-1.30
sze:grv	-0.037	0.010	-3.66
sze:orn	0.039	0.018	2.20
spt:orn	0.020	0.006	3.12
obj:orn	0.009	0.007	1.26
grv:orn	0.011	0.007	1.52
sze:spt:orn	-0.014	0.013	-1.10
sze:obj:orn	-0.003	0.014	-0.24
sze:grv:orn	-0.047	0.014	-3.25
Trial L	-0.043	0.006	-7.47
Trial Q	0.015	0.001	16.96
sze:Trial	0.018	0.011	1.60
orn:Trial	0.028	0.003	8.64
sze:TrialQ	-0.000	0.002	-0.05
orn:TrialQ	-0.006	0.005	-1.27

Table 3: Summary of the LMM fitted to the KKL data set. Extensions to the reference model are highlighted. For factors, $-0.5/+05$ dummy coding was used. L: linear, Q: quadratic.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	5.6818	0.0192	295.8684	< 0.0001
sze	0.1832	0.0384	4.7703	< 0.0001
spt	0.0744	0.0078	9.5469	< 0.0001
obj	0.0418	0.0044	9.4394	< 0.0001
grv	-0.0005	0.0050	-0.1030	0.9179
orn	0.0391	0.0157	2.4911	0.0127
sze:spt	0.0465	0.0156	2.9854	0.0028
sze:obj	-0.0095	0.0089	-1.0729	0.2833
sze:grv	-0.0379	0.0100	-3.8065	0.0001
sze:orn	0.0043	0.0314	0.1370	0.8910
spt:orn	0.0226	0.0065	3.4751	0.0005
obj:orn	0.0083	0.0070	1.1935	0.2327
grv:orn	0.0113	0.0070	1.6213	0.1050
sze:spt:orn	-0.0148	0.0130	-1.1420	0.2535
sze:obj:orn	-0.0038	0.0139	-0.2712	0.7862
sze:grv:orn	-0.0471	0.0140	-3.3658	0.0008

B. smooth terms	edf	Ref.df	F-value	p-value
fs(Trial,subj)	604.8040	772.0000	1241.7205	< 0.0001
re(subj,spt)	76.9754	84.0000	28.4032	< 0.0001
re(subj,grv)	46.1523	84.0000	1.9553	< 0.0001
re(subj,obj)	35.8780	84.0000	1.8586	< 0.0001
re(subj,orn)	54.7554	84.0000	1.9724	< 0.0001
re(subj,spt_orn)	46.8990	84.0000	1.2980	< 0.0001
s(Trial): big+cardinal	8.4515	8.6985	15.0442	< 0.0001
s(Trial): small+cardinal	8.1975	8.5267	10.9101	< 0.0001
s(Trial): big+diagonal	6.0122	6.6187	5.8248	< 0.0001
s(Trial): small+diagonal	8.1585	8.5015	11.0372	< 0.0001
s(Soa)	5.4968	6.6352	94.4320	< 0.0001

Table 4: Summary of the full GAMM fitted to the `KKL` data set. Extensions to the reference model are highlighted. For factors, $-0.5/+05$ dummy coding was used. A separate factor was defined with four levels, one for each combination of `Size` and `Orientation`, and a thin plate regression spline (`s()`) was fitted for each of its four levels. (Thin plate regression splines are smoothers constructed from a weighted sum of simple regular parametric smooth functions with a penalty for wiggliness.) `re(X,Y)` denotes random intercepts in Y for grouping factor X. The penalized factor smooth for subject (`fs(Trial, subj)` includes by-subject intercept calibration.

The baldey dataset

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	-1.03	0.06	-16.46	< 0.0001
gender=male	0.36	0.09	4.18	< 0.0001

	edf	Ref.df	F	p-value
s(LemmaFreq):gender=female	3.640	4.108	44.87	< 0.0001
s(LemmaFreq):gender=male	2.143	2.382	56.10	< 0.0001
te(word duration,Trial)	7.667	8.842	26.08	< 0.0001
re(word)	1677.386	2777.000	1.83	< 0.0001
re(word,gender)	499.984	5544.000	0.11	< 0.0001
re(subject, word duration)	18.709	19.000	68.24	< 0.0001
fs(session,subject)	163.358	198.000	97521.59	< 0.0001

Table 5: Summary of the model fit to the inverse transformed auditory lexical decision latencies in the `baldey` megastudy ($\rho = 0.2$). Factors received treatment dummy coding, `s()` denotes a thin plate regression spline, and `te()` a tensor product smooth. (A tensor product smooth constructs a wiggly surface out of restricted cubic splines with a penalty for wiggliness.) `re(X)` denotes random intercepts for grouping factor X, and `re(X,Y)` specifies random slopes for Y for grouping factor X. `fs()` denotes a penalized factor smooth.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
Intercept	-1.2176	0.0580	-20.9829	< 0.0001
gender = male	0.3861	0.0818	4.7187	< 0.0001
LemmaFreq	-0.0072	0.0006	-12.0210	< 0.0001
word duration	0.0003	0.0000	8.8784	< 0.0001
Trial	0.0267	0.0033	8.0820	< 0.0001
gendermale : LemmaFreq	0.0006	0.0006	0.9743	0.3299
word duration : Trial	-0.0000	0.0000	-6.6187	< 0.0001
B. smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
s(word)	1612.4919	2777.0000	3.6064	< 0.0001
s(gender, word)	444.6758	5544.0000	0.1059	< 0.0001
s(subject)	17.8630	18.0000	4191803.7692	< 0.0001
s(word duration, subject)	18.6988	19.0000	1439352.1318	< 0.0001
s(session, subject)	19.8883	20.0000	1576563.3632	< 0.0001

Table 6: Summary of a model for the `baldey` auditory lexical decision latencies with only linear effects.

The poems dataset

	Estimate	Std. Error	t-value	Pr(>|t|)
Intercept	6.01	0.02	334.65	< 0.0001
	edf	**Ref.df**	**F**	**p-value**
te(Fre,Trial)	10.43	11.56	79.20	< 0.0001
re(Poem)	81.08	86.00	19.40	< 0.0001
fs(Trial, Subject)	2269.74	2933.00	341.31	< 0.0001
re(Subject, Fre)	304.18	325.00	14.98	< 0.0001

Table 7: Summary of the generalized additive mixed model fitted to the `poems` data, with $\rho = 0.3$, and a tensor product smooth for `Frequency` by `Trial` (fREML 49300). `te(X,Y)` denotes a tensor product smooth, `re(X)` random intercepts for grouping factor X, `re(X,Y)` denotes random slopes for Y by grouping factor X, and `fs()` denotes a penalized factor smooth.

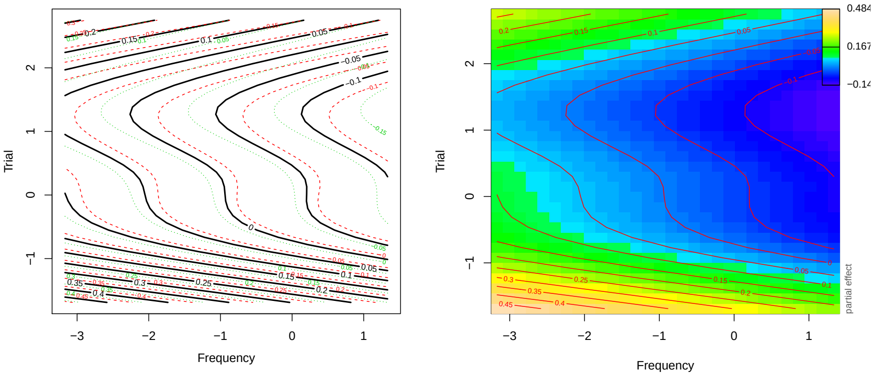

Figure 13: Tensor product smooth for the interaction of `Frequency` by `Trial` in the `poems` data set. In the left panel, green dotted lines indicate +1 standard error contour lines, and red dashed lines −1 standard error contour lines.

	Estimate	Std. Error	t-value	Pr(>|t|)
Intercept	6.04	0.02	334.97	< 0.0001
	edf	Ref.df	F	p-value
ti(Fre)	2.09	2.53	203.03	< 0.0001
ti(TrialSc)	3.93	3.94	83.34	< 0.0001
ti(Fre,TrialSc)	8.02	10.11	10.42	< 0.0001
re(Poem)	81.08	86.00	19.36	< 0.0001
fs(TrialSc, Subject)	2269.86	2933.00	337.14	< 0.0001
re(Subject, Fre)	304.16	325.00	14.98	< 0.0001

Table 8: Summary of a generalized additive mixed model fitted to the `poems` data, with additive effects of `Frequency`, `Trial`, and their interaction, and with $\rho = 0.3$ (fREML: 49307). `ti(X,Y)` denotes an independent tensor product smooth interaction term, and `ti(X)` the independent main effect of X. `re(X)` specifies random intercepts for grouping factor X, and `re(X,Y)` denotes random slopes for Y by grouping factor X. `fs()` represents a penalized factor smooth, which absorbs the by-subject random intercepts.

A. parametric coefficients	Estimate	Std. Error	*t*-value	Pr(>|t|)
Intercept	6.0390	0.0146	414.2755	< 0.0001
Fre	-0.0527	0.0020	-26.3828	< 0.0001
TrialSc	-0.0783	0.0055	-14.2661	< 0.0001
Fre:TrialSc	0.0038	0.0006	6.2019	< 0.0001

B. smooth terms	edf	Ref.df	F-value	p-value
re(Poem)	84.2246	86.0000	4097.3465	< 0.0001
re(Subject)	324.1010	325.0000	4110.0355	< 0.0001
re(Subject,Fre)	293.5805	325.0000	19.8474	< 0.0001
re(Subject,TrialSc)	319.2450	325.0000	3703.0762	< 0.0001

Table 9: Summary of a generalized additive mixed model fitted to the `poems` data, with linear terms only. `re`: random effect.