Institute for Visualization and Interactive Systems

University of Stuttgart
Universitätsstraße 38
D–70569 Stuttgart

Masterarbeit

Statistically evaluating mixed-effects models for EEG analysis using large-scale simulations

Luis Lips

Course of Study:	M.Sc. Informatik
Examiner:	Jun.-Prof. Dr. rer. nat. Benedikt Ehinger
Supervisor:	Judith Schepers, M.Sc.
Commenced:	June 15, 2022
Completed:	December 15, 2022

Abstract

Over the last decade, many EEG/ERP studies have suffered from low statistical power and bad reproducibility. Still, the adaption to alternative approaches progresses only slowly, and in most cases, researchers stick to established methods. Linear mixed effects models are a promising alternative to the established two-stage approach in the ERP analysis. However, a comparison between the two-stage approach and the linear mixed models approach in the ERP analysis concerning the number of subjects and items and the between-subject variability has not been conducted yet. In this thesis, we introduce the toolbox UnfoldSim.jl for simulating EEG data with subject and item effects. Furthermore, based on simulated ground truth data, we investigate the statistical power of the two-stage approach and the linear mixed model approach regarding the number of subjects and items as well as the between-subject variability. We observed, as expected, a gain in statistical power by increasing the number of subjects and items for both modelling schemes. In contrast to our expectation, the conducted power analysis showed no significant advantage of the linear mixed model approach against the two-stage approach concerning the number of subjects and items and varying between-subject variances. Additionally, we observed an increased type I error rate for the linear mixed model approach for simulated data with a varying subject slope. However, the results should be considered with caution. The analysis is based on simulated EEG data, and the applicability to real ERP analyses needs further investigation. The toolbox UnfoldSim.jl is a good starting point to explore the applicability to real EEG data in more detail. Furthermore, based on more realistic simulations, the UnfoldSim.jl toolbox, combined with a subsequent power analysis, could be a possibility to determine the needed number of subjects and items to reach sufficient power for future studies.

Contents

1 Introduction

Electroencephalography (EEG) is a non-invasive method for acquiring brain activity data. It is cost-efficient with a very high temporal resolution. In research, it is common to further focus on event-related potentials (ERPs) within the recorded EEG data. Event-related potentials are oscillations in the EEG signal that are time-locked to an event (e.g., stimulus). However, event-related potentials are tiny compared to other signals inside the EEG. Therefore ERPs are typically investigated by averaging over multiple trials per subject. Points of interest are then extracted for each subject and condition and further analysed. This so-called two-stage approach is common in ERP analysis. More complex approaches, like hierarchical analyses spanning all trials, have been used sparingly in the past. The biggest challenge in analysing EEG data via more complex approaches has been the time consumption and computational complexity. With computers constantly improving efficiency, traditional approaches should be questioned again, and new methods should be evaluated for improvements. One approach already being used in many areas and showing great results are linear mixed models.

Motivation & Related Work

Analysing event-related potential has led to many discoveries over the last few decades. However, more and more studies fail, trying to reproduce old findings (Pavlov et al., 2021). The Open-ScienceCollaboration (2015) attempted to replicate studies in the field of neuroscience and reported only for 36% of them similar significant findings. The reasons for the poor reproducibility are multifaceted. A significant factor are the experimental degrees of freedom in the analysis (Simmons et al., 2011). From lab to lab, a simple approach, like the two-stage approach, can be applied in different variations. Researchers often rely on long-standing or established analysis techniques in their lab despite already-known flaws (Boudewyn et al., 2018). A direct consequence is very low statistical power of studies. Button et al. (2013) already reported that many studies in neuroscience yield very low statistical power (8% - 30%). Lower statistical power decreases the chance of reproducibility and significant results.

A simple approach to improve statistical power is to increase the investigated sample size (e.g. subjects examined). In practice, this is often inapplicable. The number of trials presented per study is a frequently overlooked possibility to improve power(Boudewyn et al., 2018). Baker et al. (2021)

investigated the effects of the number of trials per subject for multiple experimental paradigms and methodologies, including EEG and fMRI. They showed for multiple cases that statistical power could be maintained with a smaller subject size but increased trial number per subject.

Linear mixed models seem to be another possibility to improve the statistical power in various scenarios. For the use of LMMs in hypothesis testing, a wide range of related work has already been conducted. Barr et al. (2013) investigated the effects of different model specifications on the type I error. They compared random-intercepts-only models and models with random slopes regarding the type I error and yielded statistical power. Barr et al. (2013) recommend using the maximal random effects structure to minimise the type I error.

The study by Matuschek et al. (2017) builds on the findings of Barr et al. (2013) and states that using the maximal random effect structure increases the type II error and hence lowers the statistical power. Researchers using models with maximal random effect structures in studies should thus be aware of the consequences.

LMMs have also been applied to EEG analysis. Because of the cost-efficiency, the robustness advantage over other methods, and the high temporal resolution, EEG is used in a very broad area. Not always, perfect experimental conditions are guaranteed. In their work, Heise et al. (2022) inspected the consequences of unbalanced designs (designs with unequal observations per subject) on statistical models. The paper focuses on EEG research in infants and children where it is challenging to collect data due to practical problems like, e.g., short attention or excessive movement. They compare traditional two-level models to linear mixed models for an unbalanced design. They propose linear mixed-effects models (LMMs) to mitigate the generated bias of an unbalanced design in the traditional two-level models.

With EEG/ERP research evolving increasingly towards more complicated experiment designs, this leads to problems with established statistical analysis methods and approaches (e.g., two-way approach and summary statistics). In particular standard ways of group analysis, like the two-stage approach, do not incorporate within-subject uncertainty and have difficulties with unbalanced designs. As a consequence, the result can be biased or even false (Heise et al., 2022). Linear mixed-effects models offer multiple advantages over the established two-level approach. In particular, with a focus on ERP analysis, there needs to be more research on how the two-stage approach compares to the linear mixed model analysis in various scenarios.

The thesis intends to investigate the effect of a varying number of subjects, the number of items per subject, and the between-subject variability on the statistical power with regard to the selected modelling schemes. The comparison is conducted on simulated EEG data to control for the mentioned parameters. Therefore, the thesis is separated into three main goals. The first goal is the creation of a simulation toolbox to create ground-truth EEG simulations. The second goal is the implementation of the different modelling schemes, and the third is conducting the comparison concerning the parameters as mentioned above.

2 Foundations

The intent of this chapter is not to be an in-depth explanation of each topic. It serves the purpose of an introduction, a reminder of the terminology, and fundamental support to the reader.

Electroencephalography

Electroencephalography, short EEG, is a non-invasive technique to measure brain activity. Therefore small electrodes are placed on the scalp. There are multiple patterns to place the electrodes. A typical system is the international 10-20 system (Luck, 2014). The electrodes amplify the underlying source potentials and capture them in the EEG signal. This technique has natural limitations. EEG can measure not every brain potential inside the brain. Between the potential and the electrodes is the human skull. Potentials created by outside layers of the brain are favoured. Potentials of deeper sources in the brain are less likely to be measured by EEG. Another limitation is the orientation of the source potentials. Potentials parallel to the skull are challenging to measure via EEG. Opposing potentials can also cancel each other out. A direct consequence of those natural and physical limitations is low spatial resolution (Jackson and Bolger, 2014). In contrast, however, the temporal resolution of EEG measurements is very high. Hence EEG is often used in experiments measuring the response time to specific events. In comparison, fMRI, a method of measuring brain activity based on blood flow in the brain, has a lower temporal resolution but a very high spatial resolution.

Event-related potentials

Hidden within an EEG measurement are so-called event-related potentials (ERPs). Event-related potentials describe the response of our brain (voltage fluctuations) time-locked to an event. An event can be visual or auditory – motor or cognitive events are also possible. A simple example is presenting a visual stimulus like a picture of a car or face. ERPs are not visible in the raw EEG signal because of their small voltage compared to the overall EEG signal. A widespread technique in analyzing ERPs is to average over multiple trials to improve the signal-to-noise ratio. Different

events trigger different responses. Typically event-related potentials of multiple conditions are compared against each other. The ERP analysis aims to draw conclusions about brain functions based on the detected differences between the event-related potentials (Luck, 2014).

The voltage of an ERP ranges between -20 to 20 μV. The epoch window (timespan of interests) can span up to 1000ms and varies for each ERP. Exact values can change depending on the ERP in focus. As a comparison, spontaneous fluctuations in EEG can reach 10 to 100 μV. Muscular artifacts even up to 1000μV (Kappenman et al., 2021).

Experiment design

Many ERP experiments focus on detecting differences between two or more conditions. Conditions, as before, can be different visual stimuli, for example, images of faces and cars. Current studies primarily investigate effects under optimal conditions inside a lab. Nevertheless, more complex and naturalistic experiment designs are becoming more common. Subjects are confronted with multiple stimuli during an EEG recording to capture the individual event-related potentials. The experiment design specifies the conditions each subject is exposed to. We discriminate between a within-subject design and a between-subject design.

Between-subject design

A between-subject design exposes each participant only to one condition.
Example: Consider a simple experiment with only two subjects, two items and two conditions. In this case, each item corresponds to one condition. Subject 1 is only exposed to condition A, and subject 2 is only to condition B.

Subject	Item	Condition
S1	I1	A
S2	I2	B

Table 2.1: Between-subject design

Within-subject design

The within-subject design, in contrast, exposes each participant to every condition.
Example: Consider the same simple experiment with only two subjects, two items, and two conditions. In the within-subject design, each subject is exposed to every condition. Subjects 1 and 2 each see items 1 and 2 (respective conditions A and B).

Subject	Item	Condition
S1	I1	A
S1	I2	B
S2	I1	A
S2	I2	B

Table 2.2: Within-subject design

It is important to mention that sometimes the independent variable dictates the experiment design (e.g. age). Experiment designs can also be combined if required.

Statistics

Hypothesis testing

ERP analysis often aims to make statements about the difference between conditions. A problem caused by this is to differentiate between significant effects and effects that only occurred by chance. Inferential statistics, more specific hypothesis testing, is an entire field dedicated to this question. Hypothesis testing distinguishes between the null hypothesis and the alternative hypothesis. The null hypothesis states that all the data comes from the same distribution. An effect between conditions is assumed to be absent. The alternative hypothesis acts as a counterpart to the null hypothesis. It states that the inspected data comes from different underlying distributions. By default, we assume the null hypothesis generates the data. The alternative hypothesis is only accepted if the probability that the null hypothesis generated the data is below a small threshold (Daniel and Cross, 2013).

Type I and Type II eror

When using hypothesis testing to infer about the underlying distributions, by chance, the wrong inference can be made. An error is classified as a type I error or type II error.

Null hypothesis is...	**true**	**false**
rejected	Type I error α	True positive $1-\beta$
accepted	True negative $1-\alpha$	Type II error β

Table 2.3: Type I & Type II Error

A decision is classified as a type I error if the null hypothesis is true but rejected. In other words, we state that there is a significant effect in the case of no real effect. A type II error, or false negative, happens if the null hypothesis is false but not rejected. To rephrase it again, a type II error happens if an existing effect is not detected (Daniel and Cross, 2013).

p-value

P-value is short for probability value. The p-value is the probability that the data materialized from the null hypothesis (Daniel and Cross, 2013). The smaller the p-value, the smaller the chance that the null hypothesis generated the data. Vice versa, a high p-value corresponds to a high probability that the null hypothesis generated the data. A p-value is typically the output of a statistical test (e.g. t-test, permutation test). Together with the significance level, it helps to determine if the null hypothesis can be rejected. The significance level α defines the accepted margin of error. A significance level of $\alpha = 0.05$ corresponds to a 5% chance of a type I error. Simply put, in 5% of the cases, the statistical test outputs a significant p-value below 0.05 when there is no effect in the data.

Statistical power

Statistical power is the probability of correctly rejecting the null hypothesis. In other words, power is the probability of a true positive. It reflects the probability of finding an effect between, e.g. conditions if there is one. The statistical power $(1 - \beta)$ is directly linked with the type II error rate (β) (Daniel and Cross, 2013).

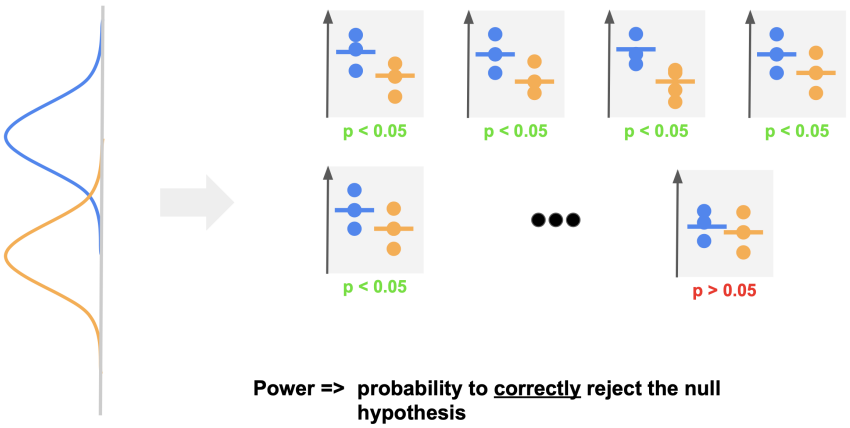


Figure 2.1: Statistical Power

The computation is typically repeated for different data samples, as shown in Figure 2.1, to improve the estimate of the statistical power. Each grey box corresponds to a new set of samples. The blue and orange dots are samples drawn from different distributions. The estimated power corresponds to the fraction of significant p-values with respect to the number of iterations.

3 Research Question

The purpose of this master thesis is to investigate the influence of a varying number of subjects, a varying number of trials per subject, and the between-subject variability on the statistical power of detecting an event with regard to selected modelling schemes. The modelling schemes selected for the comparison are the two-stage approach and linear mixed models.

The comparison is conducted on simulated data. Due to the multiple variable parameters, the experiments are limited to a subset of all combinations. The research focuses on a simple within-subject design with two conditions. All experiments are simulated with up to 50 subjects. The same applies to the number of items. Between-subject variability is inspected via subject effects in the form of random intercepts and slopes. The measure for the comparison of the modelling schemes is the statistical power. Power contour plots help to visualize the statistical power for different numbers of subjects and items. Figure 3.1 depicts this process from left to right.

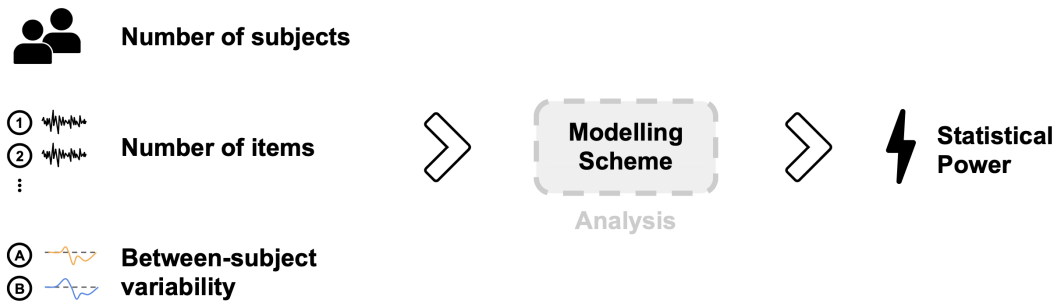


Figure 3.1: Research question

We expect that more sophisticated models, like linear mixed-effects models, are more flexible and suitable for modern EEG analysis. We predict that the statistical power is higher for the linear mixed model approach, especially in cases where only a small number of subjects and trials per subject are available. If so, LMMs should be preferred over the two-stage approach to achieve a desirable statistical power above 80% for more reproducible studies in neuroscience.

4 Methods

As stated in the previous chapter (Chapter 3), this thesis aims to provide a starting point to investigate and compare different ERP analysis methods. Since no ground truth concerning the effect size and between-subject variability is available for measured EEG data, EEG-like data is simulated for selected parameters. The following chapter seeks to provide a more detailed look at the methods. Methodologically, the thesis is split into two parts — the simulation of data and the subsequent analysis.

4.1 Simulation via linear mixed model parameterization

The goal is to simulate EEG-like data to compare ERP analysis methods. Accordingly, ERPs for different conditions and parameters need to be simulated. In real ERP studies, the response to an event varies from subject to subject. Each subject has a different baseline and is susceptible to conditions in different ways. The between-subject variance is a measurement of how much subjects differ from one another. The simulation is implemented as a linear mixed model parameterization to control the between-subject variance. For simplicity, data is only simulated for a single channel.

It is important to mention that the simulation does not intend to model the EEG signal based on the real underlying source space. It aims to provide ground-truth data regarding different effect sizes and between-subject variances. The data is generated in the following way:

1) Specifying design and parameters

   The fixed and random effect structure is defined via the model formula in Wilkinson notation (Wilkinson and Rogers, 1973). In the experiments, data is simulated with a categorical fixed effect with two levels. The analysis was fixed to within-subject designs. Each subject was therefore exposed to each condition. Additional to the fixed effects, different combinations of subject effects are included in the simulated dataset. ERPs were simulated from the following underlying model:

$$y \sim 1 + \text{cond} + (1 + \text{cond} \mid \text{subj})$$

2) Sampling

   The model defines a random intercept and a random slope per subject. In each simulation run, the variance of the distributions of the random effects is fixed to the specified value. The random intercept and slope per subject are then sampled according to the distribution. Additional random variation per subject-item combination is added based on the defined within-subject variance.

3) Expanding over time

   In order to move from individual samples per subject-item combination to a simulated ERP over time, the sampled value is multiplied by a basis function. The basis function defines the waveform of the ERP over time.

4) Sampling onsets without overlap

   Event onsets are sampled such that ERPs do not overlap. Therefore an offset between different event onsets with some random jitter is created. The onsets alternate, within every subject, between each condition level. Each condition is equally represented in each subject (balanced design).

5) Convolve onsets and ERPs to EEG signal

   The last step is the convolution of the beforehand created ERPs and onsets. On top of the simulated EEG signal, additional noise is added. The result is a continuous simulated EEG signal for the specified number of subjects and items with a known magnitude of effect size and between-subject variance.

4.2 Analysis pipeline

The thesis aims to compare different modelling schemes concerning the statistical power in detecting an effect. As described before, the two-stage approach and linear mixed models are selected for comparison. The following chapter describes the individual steps of the analysis in more detail. The analysis pipeline is separated into four main parts (Figure 4.1). The input and starting point for the analysis is the simulated EEG data. The simulation does not include prominent noise sources like muscle artifacts or 50/60 Hz noise. Therefore, the first preprocessing steps of an ERP analysis on real data, cleaning and filtering, are omitted.

Figure 4.1: Analysis pipeline

The simulated data mimics a continuous EEG recording over time. In the first step the continuous data is split into time-locked epochs. The output is a single epoch for each subject-item combination.

In the next step, the data is further condensed by averaging each epoch over a time window around the peak. The averaging is more robust than only extracting the peak. It also bypasses the multiple-comparison problem by eliminating the time dimension — the averaging results in a single value per subject-item combination.

Those values per subject-item combination are then passed to the modelling scheme and the subsequent statistical test to test for significance by computing the p-value.

4.3 Modelling schemes

4.3.1 Two-stage approach

This approach separates the analysis into the individual subject level and the group level. The individual subject level contains the trials per subject (for multiple conditions). The group level spans over all subjects. In the two-stage approach, the trials on the individual subject level are condensed into a single value per subject-condition combination. This is achieved by fitting a simple linear regression to each subject. In this case, this is similar to averaging over trials grouped by subject and condition. Smith and Kutas (2015) showed the similarity by computing the least squares solution to the simple linear regression of the form $y_i = \beta + \epsilon$. The least-square solution is:

$$\frac{1}{n}\sum_{i=1}^{n} y_i = \beta$$

This is equal to computing the mean over trials per subject-condition combination. In the second stage, the extracted values (averaged values) are then further analyzed using, e.g. a paired t-test. Based on the p-value, a conclusion can then be drawn about the significance of the overall effect on the population.

The Figure 4.2 depicts the two-stage approach schematically. Input for the analysis are single trials. Each trial is visualized via a circle. The colour shows the belonging to the subject. Filled and unfilled circles visualize conditions. A model per subject is fitted on the first level, described

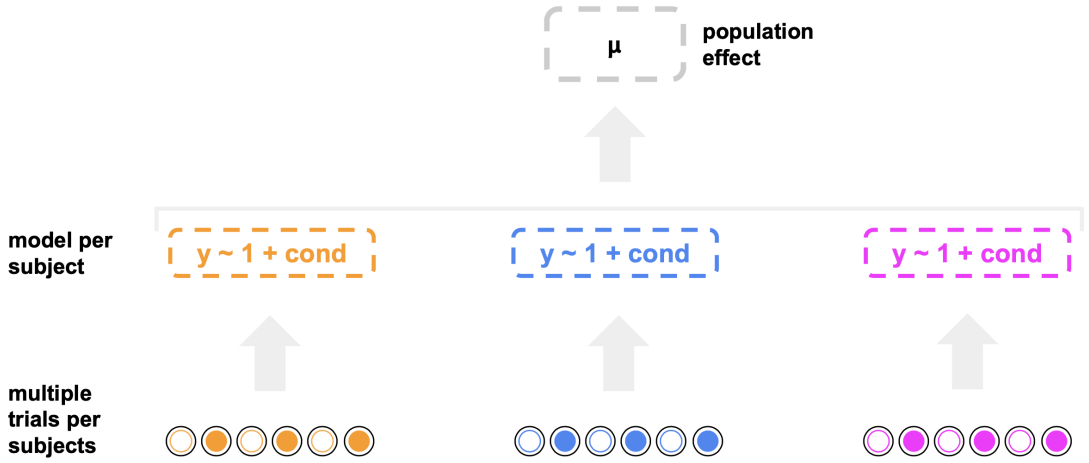


Figure 4.2: Two-stage approach

as the individual subject level before. This is visualized by the arrow pointing to the formula in Wilkinson's notation. Resulting values are then used to further generalize about the population effect.

This procedure simplifies the complexity of the analysis. On each level, the dimension is further broken down by aggregating data. This is reflected in the fast computation time. An intended side effect of averaging over multiple repetitive trials per subject is the increase of the statistical power and elimination of noise. Nevertheless, this also has its drawbacks and limitations. One major drawback is the loss of information about the uncertainty of the extracted location parameter. A more elaborate modelling scheme are linear mixed-effects models, which explicitly take into account uncertainties at multiple levels.

4.3.2 Linear mixed model approach

Linear mixed-effects models (LMMs) have a long-reaching history. Since their first application to psycholinguistic data (Baayen et al., 2008), LMMs gained more and more popularity in psycholinguistics and other research fields. Especially where nested experimental data or repeated measurements are encountered. Consequently, they come in different shapes and names. Linear mixed-effects models are also called nested data models, multilevel models, or hierarchical models. LMMs are an extension of simple linear models. Unlike simple linear models, linear mixed-effects models can describe fixed effects (average effect of the independent variable on a dependent measure) and random effects (random variation from subject to subject or item to item). Laird and Ware (1982) defined linear mixed-effects models the following way:

$$y = X\beta + Zu + \epsilon$$

$$u \sim N(0, G)$$

$$\epsilon \sim N(0, R)$$

The vector y represents the observations. In the context of EEG, this corresponds to the recorded EEG data. β is an unknown vector of fixed effects. u is an unknown vector of random effects with mean $E(u) = 0$ and variance-covariance matrix $var(u) = G$. X and Z are design matrices for the respective effects (fixed / random). ϵ is a vector of random errors, with mean $E(\epsilon) = 0$ and variance $var(\epsilon) = R$. This allows the possibility to model parameters on the group level and within each subject. A common method to compute the parameters is restricted maximum likelihood estimation.

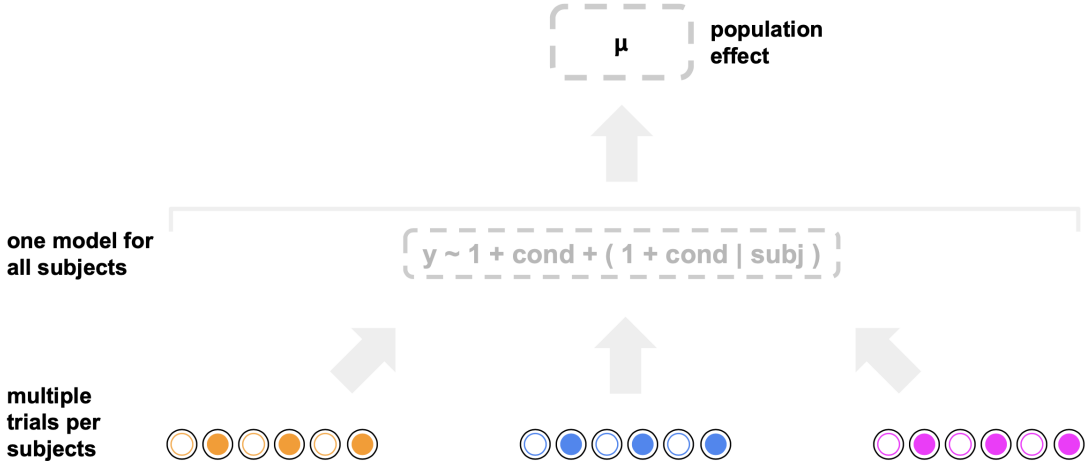


Figure 4.3: Linear mixed model approach

The figure above visualizes the application of LMMs in the EEG/ERP analysis. The lower input level is similar to the visualization of the two-stage approach. Each dot represents a single trial. Belonging to a condition is visualized by filled and unfilled dots. The colour determines the subject the trial belongs to. The figure shows three subjects (orange, blue, and pink) with six trials each (3 per condition). All the trials are then passed to a single linear mixed model of the form:

$$\text{y} \sim 1 + \text{cond} + (1 + \text{cond} \mid \text{subj})$$

This model has a maximal random effect structure (Barr et al., 2013). Besides the fixed effects intercept and condition, it models a random intercept and slope per subject. The fitted model is then evaluated for significance. The significance testing of linear mixed models is still an open

discussion with no gold standard available. The possibilities range from likelihood ratio test to permutation test and many more. The most significance tests for LMMs suffer from a high type I error (Barr et al., 2013; Matuschek et al., 2017; Warrington et al., 2014).

5 Implementation

The entire pipeline, the simulation and the analysis, is implemented in the programming language Julia. The implemented procedure, independent of the modelling scheme, is split into two parts. The first part comprises the data simulation and the subsequent computation of the p-value. The second part is the computation of the statistical power. Figure 5.1 below showcases the procedure schematically.

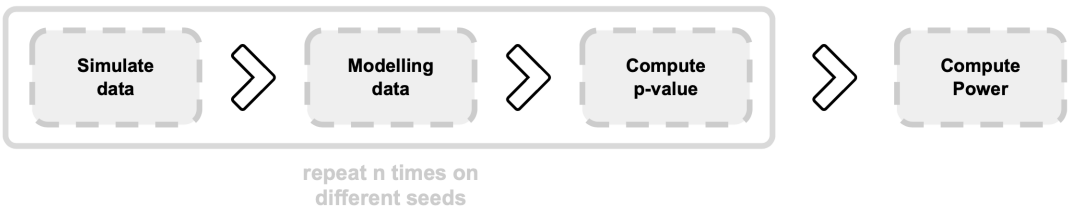


Figure 5.1: Implementation overview

The process of simulating data and computing the p-value is thereby repeated multiple times (1000), to achieve a more confident estimate of the statistical power. In each iteration, completely new data is simulated by changing the seed.

5.1 Simulating data

To simulate data according to the described linear mixed model parameterization in the chapter Methods, the toolbox UnfoldSim.jl was created. UnfoldSim.jl provides a simple and coherent way to simulate ERPs and continuous EEG signals. It allows the simulation for different experiment designs and varying number of subjects and items. Also, different ERP components can be imitated based on customizable and combinable basis functions. For each component, different effect structures can be defined. The code and current state of the package UnfoldSim.jl is visible under `https://github.com/unfoldtoolbox/UnfoldSim.jl`. Furthermore, a quick introduction and example is provided in the corresponding `README.md`.

5.1.1 UnfoldSim.jl

A simulation begins with defining the corresponding experiment design and components of the simulation. The experiment design specifies the number of subjects and items and the within- or between-subject factors (or both). A component in the simulation represents a real ERP component. Each component consists of a basis function, a model formula, the effect size beta, the random effect variances, and the residual variance.

Basis function
The basis function describes the waveform of the component over time. For each time sample, the function defines a value between -1.0 and 1.0. The length of the basis function (in samples) defines the length of the component. A simple basis function imitating a peak is a Hanning window.

Formula
The next and essential part is the formula. The formula is specified in Wilkinson's notation. The formula is important since we sample random effects from a linear mixed model. Hence, the formula specifies the fixed and random effects structure. Below are two examples of a formula in Wilkinson notation with random effects:

$$\mathrm{y} \sim 1 + \mathrm{cond} + (1 + \mathrm{cond} \mid \mathrm{subj}) \tag{5.1}$$

Based on Equation (5.1) the toolbox would sample a random intercept per subject and a random slope for condition per subject. Alternative also item effects can be included in the simulation. The formula shown in Equation (5.2) samples additionally a random intercept per item.

$$\mathrm{y} \sim 1 + \mathrm{cond} + (1 + \mathrm{cond} \mid \mathrm{subj}) + (1 \mid \mathrm{item}) \tag{5.2}$$

Effect size
The parameter β specifies the effect sizes of the fixed effects. The dimension is implicitly given by the formula. For Equations (5.1) and (5.2)), β has the dimension 1x2. The first entry corresponds to the intercept and the second value to the effect size of cond.

Random effect variance
The between-subject variance is defined via σ_{ranef}. It specifies the variance of the random effect distributions. The random effects structure and dimension is defined by the formula.

$$\mathrm{y} \sim 1 + \mathrm{cond} + (1 + \mathrm{cond} \mid \mathrm{subj}) \longrightarrow \text{variance for subject intercept \& slope}$$
$$\mathrm{y} \sim 1 + \mathrm{cond} + (1 \mid \mathrm{subj}) + (1 \mid \mathrm{item}) \longrightarrow \text{variance for subject \& item intercept}$$

Residual variance
σ_{res} corresponds to the residual variance of the linear mixed model.

Noise

In the current implementation UnfoldSim.jl also provides the possibility to add noise to the simulation. The noise is added after the convolution of the ERPs and event onsets to a continuous EEG signal. Currently, white noise, red noise and pink noise are supported and can be specified via *noisetype*. The extent of added noise can be adapted with the variable *noiselevel*.

5.1.2 Generated data for comparison

For the comparison of the modelling schemes, ERPs are simulated based on a Hanning window basis function. The resulting simulated ERP component length in samples is 125. Figure 5.2 demonstrates the used basis function.

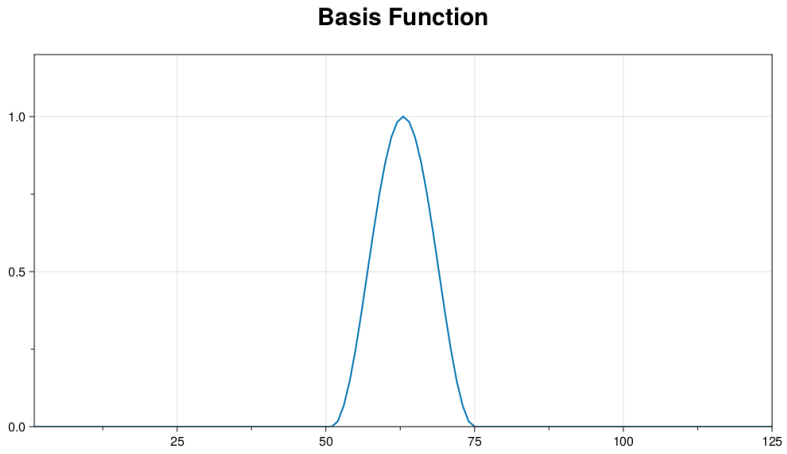


Figure 5.2: Hanning window as basis function

The sampling of different random effects is implemented using MixedModels.jl (Bates et al., 2022) and MixedModelsSim.jl (Alday, Bates, et al., 2022). Therefore a mixed model is used in reverse. Typically mixed models are used to estimate the underlying distributions of fixed and random effects. Here we specify a mixed model, fix the underlying distributions and sample from it. The resulting values, one per subject-item combination, are expanded over time via the basis function. Figure 5.3 shows simulated event-related potentials for two subjects and two items. The ERPs in Figure 5.3 were sampled with an effect size of 0.5 as well as random subject intercept variance and random subject slope variance.

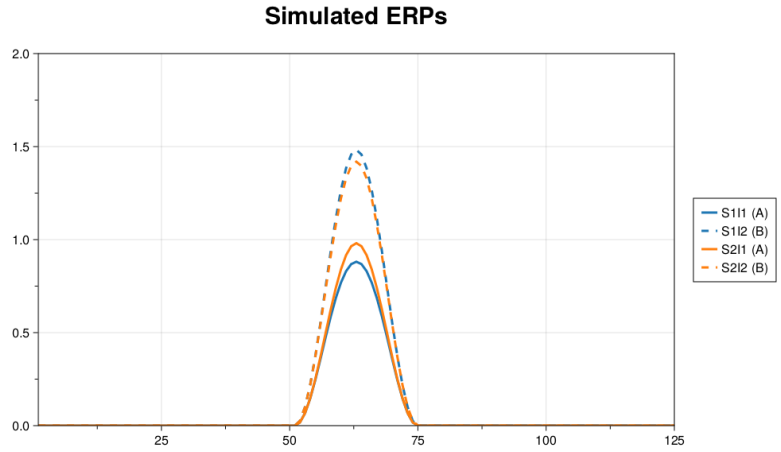


Figure 5.3: Simulated ERPs (nsubj=2, nitem=2)

Figure 5.4 shows the end result of a simulation, the continuous EEG signal. In this example, besides the two items before, four more items per subject are simulated and integrated into the EEG signal. The outcome is an over 3000 sample long EEG signal for two subjects. To expand the simulated ERPs (Figure 5.3) to an EEG signal (Figure 5.4), additional information about the event onsets is needed. In the current implementation, event onsets are created based on an average distance of 305 samples with some jitter around them. The distance is set large enough to avoid overlaps. Furthermore, pink noise is added on top of the simulated EEG signal.

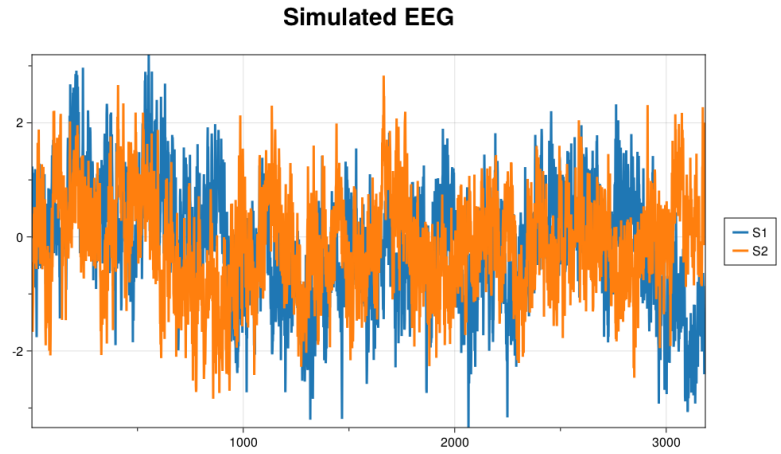


Figure 5.4: Simulated continuous (noisy) EEG signal (nsubj=2, nitem=6).

5.2 Computing p-values

The continuous EEG signals were split into separate epochs via the toolbox Unfold.jl (B. V. Ehinger and Dimigen, 2019). The selected epoch window was -0.1 to 1.2 seconds. Time-independent values are created by averaging over 30% of the samples of the peak. Afterwards, the values are passed to the respective selected modelling scheme (Figure 5.5).

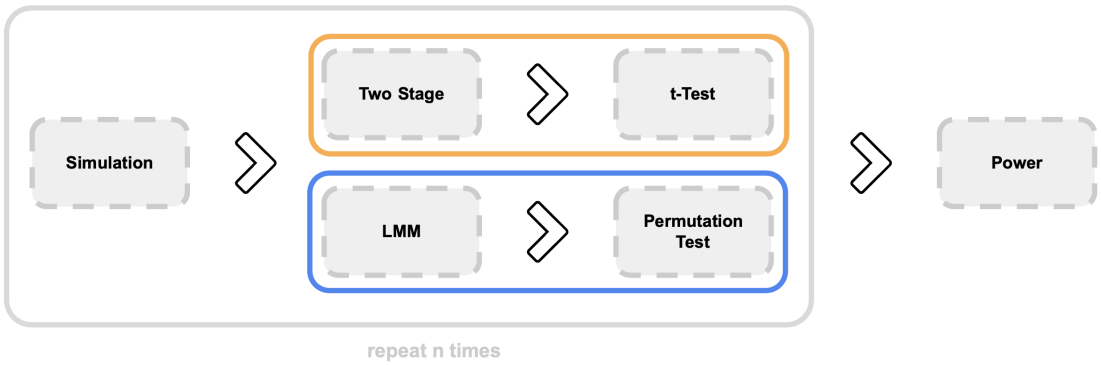


Figure 5.5: Computation p-value

5.2.1 Two-stage approach

The two-stage approach is implemented by simply averaging over the epoched data grouped by subject and condition. The alternative solution, fitting a separate simple regression model per subject is implemented but not used. The p-value is computed via a paired t-test. For the t-test, the package HypothesisTests.jl is used (*HypothesisTests.jl* 2022).

5.2.2 Linear mixed model approach

The linear mixed model approach is implemented using MixedModels.jl (Bates et al., 2022). For significance testing a permutation test provided by MixedModelsPermutation.jl is implemented (Alday, B. Ehinger, et al., 2022). The used analysis model is specified by the formula `y ~ 1+cond+(1+cond|subj)`. Hence, a random intercept and slope is estimated for each subject. Barr et al. (2013) proposed to use a maximal random effects structure for LMMs to minimize the type I error. We followed this approach, aware of the consequences of an increased type II error rate and thus lower statistical power reported by Matuschek et al. (2017). As a fast alternative, because of the time complexity of the permutation test, the by MixedModels.jl provided p-value estimate (Z-test) is used. However, the advantage of speed comes with some disadvantages as we later see in the section Chapter 6. For the comparison, the permutation test was used for significance testing in the LMM approach.

5.3 Computing power

Input to the power computation are the p-values from the respective model. The statistical power is estimated by computing the fraction of p-values below the significance level alpha. The significance level alpha was set to 0.05. The power is computed on 1000 p-values for each subject-item combination.

5.4 Computing type I error

Type I errors, or so-called false positives, determine how precisely we can compare the modelling schemes with each other. A perfect method maintains a steady type I error rate around the significance level alpha. However, the type I error can be affected by numerous factors. Besides the modelling scheme, one major factor is the hypothesis test (Matuschek et al., 2017). The type I error is computed as a sanity check for each model to determine deviations from the expected type I error of 5% for each model. The type I error is computed by simulating data with no effect via the toolbox UnfoldSim.jl and subsequently calculating the power.

6 Results

6.1 Type I Error

The following section compares the observed type I error for the paired t-test (two-stage approach) and the permutation tests (LMMs) based on simulated data. Furthermore, a comparison between the via MixedModelsPermutation.jl computed p-value and the Z-test estimate provided by MixedModels.jl is done. The significance level for all hypothesis tests is set to $\alpha = 0.05$. The expected type I error accordingly should be around 5%.

For each model, we look first at the type I for increasing subject intercept variance and subsequently for increasing subject slope variance. Each subplot visualizes the type I error for a different variance. The variance increases from left to right and top to bottom. Individually lines correspond to a different number of subjects. Light colour corresponds to a low number of subjects, and darker colour to a higher number of subjects. The number of items is shown on the x-axis. The y-axis is the type I error percentage.

6.1.1 Two-stage approach

Figure 6.1 shows the type I error for varying subject intercept variance of size 1.0, 2.0 and 3.0. The type I error does not vary extensively for different numbers of items besides slight random variation. There is no effect on the type I error visible for increasing subject intercept variance. The colours representing the numbers of subjects seem to be also randomly permuted, indicating no effect.

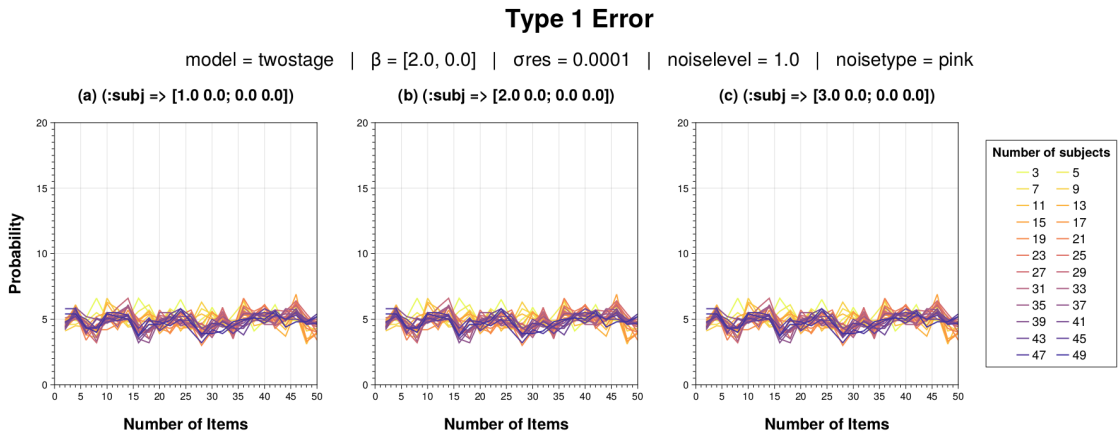


Figure 6.1: Type I error of the two-stage approach with random subject intercept variance

Figure 6.2 visualizes the type I error in the same way, but for an increasing subject slope variance. Similar to the increasing subject intercept variance, the type I error does not deviate much from the expected 5%. For an increasing subject slope variance, it is visible that the variation within the type I error increases slightly with the number of items. Overall, the type I error of the two-stage approach is calibrated and fluctuates around 5%.

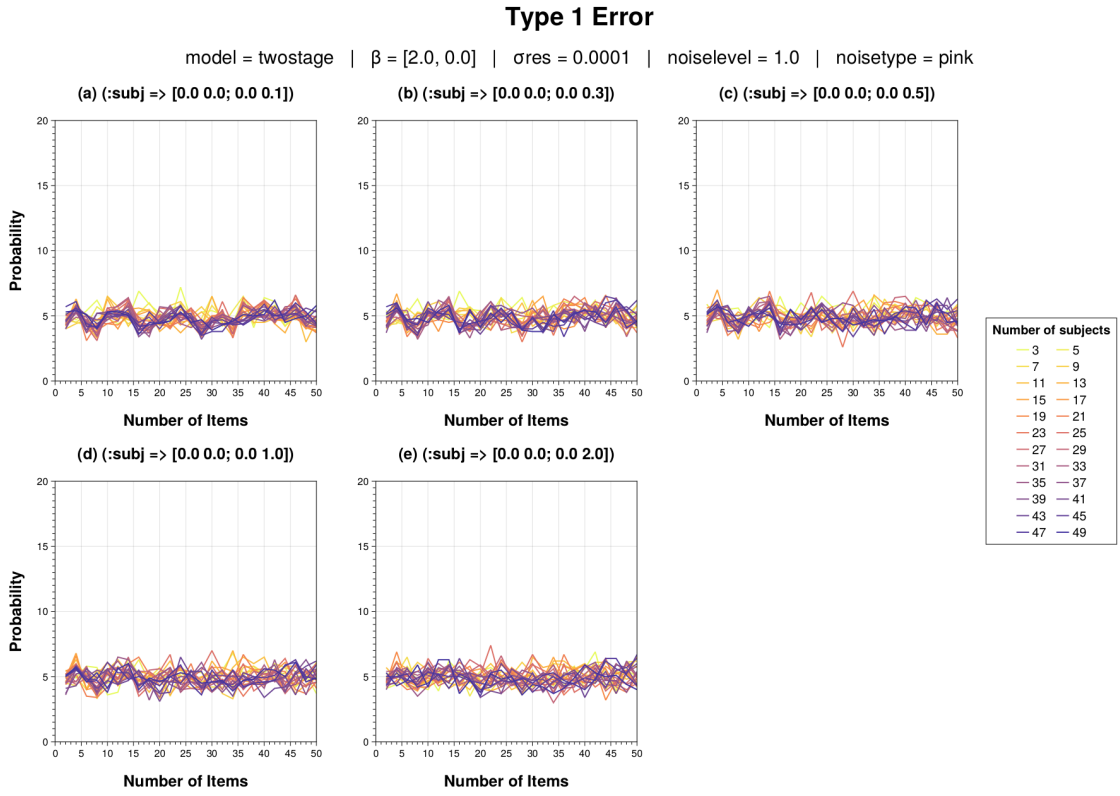


Figure 6.2: Type I error of the two-stage approach with random subject slope variance

6.1.2 Linear mixed model approach

The following section looks at the type I error for the linear mixed model approach in combination with the permutation test. Figure 6.3 shows the type I error for different increasing subject intercept variances (1.0, 2.0, 3.0). The plots show no effect of the varying subject intercept variance. Additionally, despite some outliers, the type I error is for nearly all combinations below 5%. There is no effect of the number of items or number of subjects on the type I error recognizable.

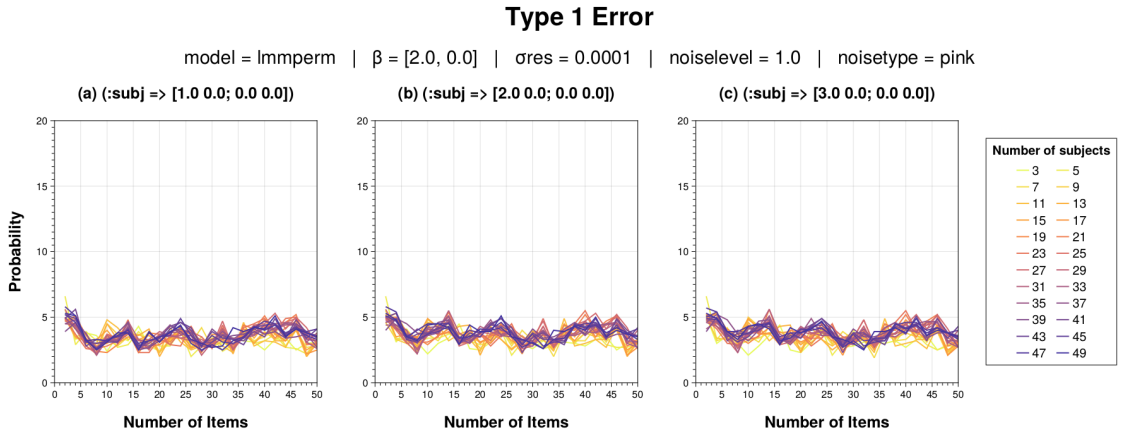


Figure 6.3: Type I error of the LMM approach with random subject intercept variance

Figure 6.4 visualizes the type I error of the linear mixed model approach for an increasing subject slope variance. For a variance of 0.1 (Figure 6.4a), the type I error is conservative. Only some outliers exceed above 5%. With an increase of the variance, it is analogous to the t-test visible that with a growing number of items, the type I error variates more. However, it is far more affecting the LMM approach. For a variance of 0.3 (Figure 6.4b), the type I error is around 5%, and for 0.5 (Figure 6.4c) already above 5% for some subjects. It is visible that for a smaller number of subjects, the type I error increases with a growing number of items. In particular, an experiment design with three and five subjects stands out (green & yellow line). For both, the type I error extends far above the 5%. With three subjects, the type I error reaches above 10%. For a subject slope variance of 1.0 (Figure 6.4d), this effect is evident for all subjects. However, the type I error seems to decrease again for experiment designs with more than 25 items. Still, the type I error rate exceeds 5% by a lot. The type I error of the permutation test of linear mixed models, thus, is for a subject slope variance equal to 1.0 too high and not calibrated. For a variance of 2.0 (Figure 6.4e) this effect is not detectable. Only for experiment designs with a low number of subjects (three and five), the type I error is again significantly increased.

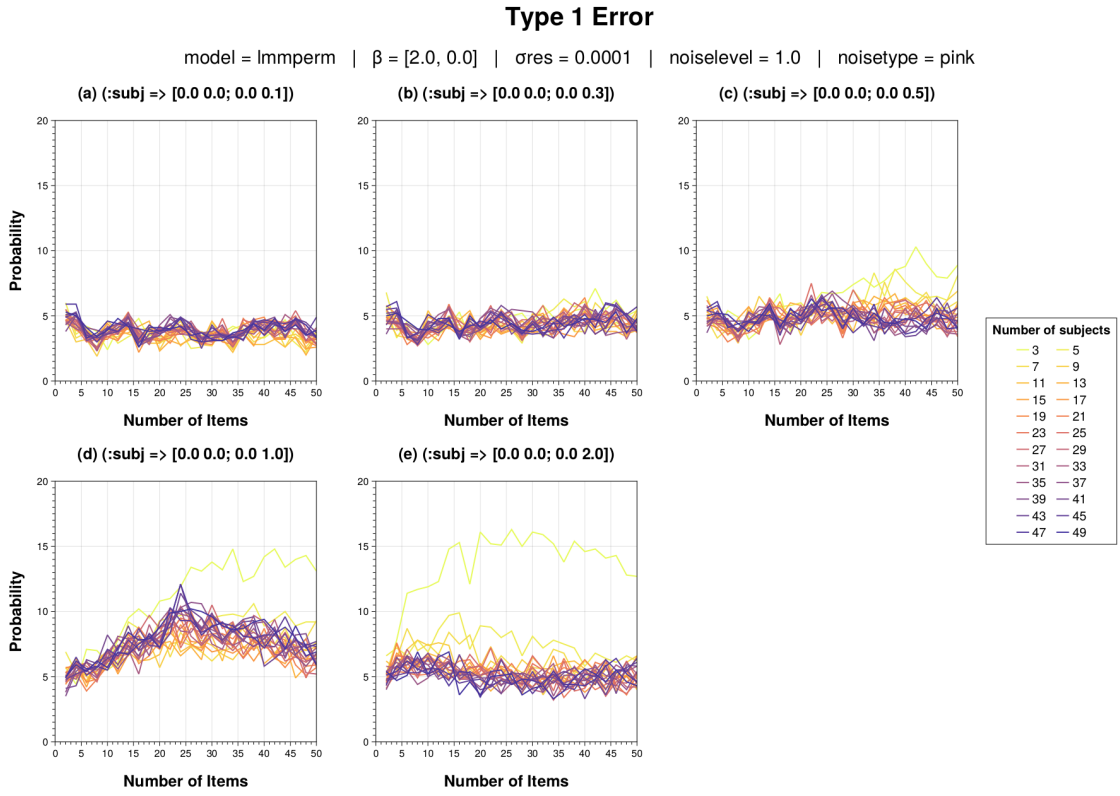


Figure 6.4: Type I error of the LMM approach with random subject slope variance

6.1.3 LMM approach - permutation test versus Z-test

As in the Chapter 5 mentioned, the implementation includes different methods to compute the p-value for linear mixed models — the Z-test provided by the MixedModels.jl package and the permutation test provided by MixedModelsPermuation.jl. In the following, we will look at the type I error for the Z-test and compare it to the previously evaluated permutation test. For the the subject intercept variance, the Z-test shows similar to the permutation test, a type I error rate below 5%. However, Figure 6.5 displays an increased type I error rate of the Z-test for experiment designs with a small number of subjects and a small number of items. The less subjects are used, the higher is the type I error. In comparison, the permutation test did not show such extreme outliers for small experiment designs. For a higher number of items (> 10), the type I error decreases to below 5% and does not show any further outliers.

The Figure 6.6 visualizes the type I error of the Z-test for different random subject slope variances. The subplots a-e show the variances 0.1, 0.3, 0.5, 1.0 and 2.0. For all variances, likewise to the subject intercept variance, an increased type I error for a small number of subjects (< 11) and items (< 4) is observed. With only 0.1 variance of the random subject slope (Figure 6.6a), the type I error is, without consideration of the outliers for a small number of subjects and items, below 5%.

For an increasing subject slope variance (0.3, 0.5, 1.0), the type I error rate for the Z-test shows a similar effect as observed for the permutation test. The higher the variance, the more the type I error increases. However, the increased type I error is far more severe visible for the Z-test. By the ordering of the colours, it is visible that smaller subject numbers are more affected than higher ones. For a subject slope variance of 2.0, the effect of an increased type I error lessens for experiment designs with a high number of subjects. For experiment designs with a small number of subjects the type I error increases even more.

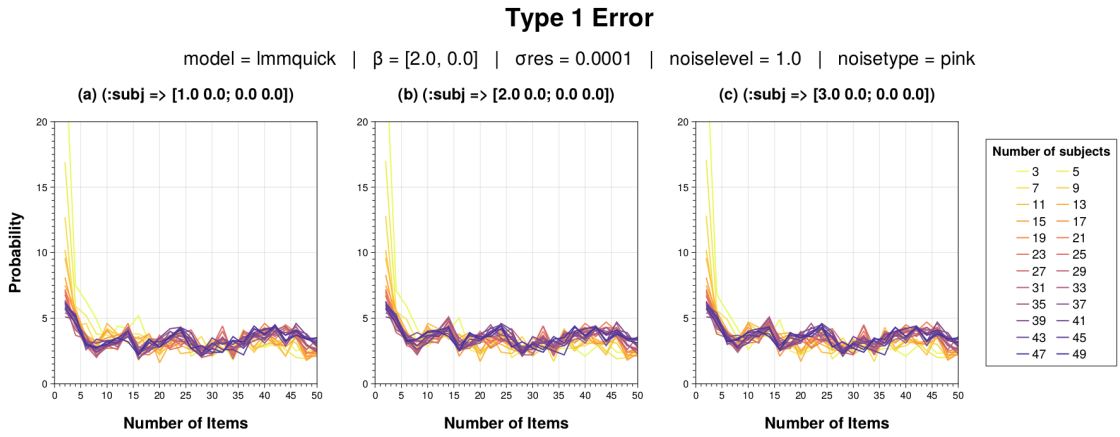


Figure 6.5: Type I error of the LMM (Z-test) with random subject intercept variance

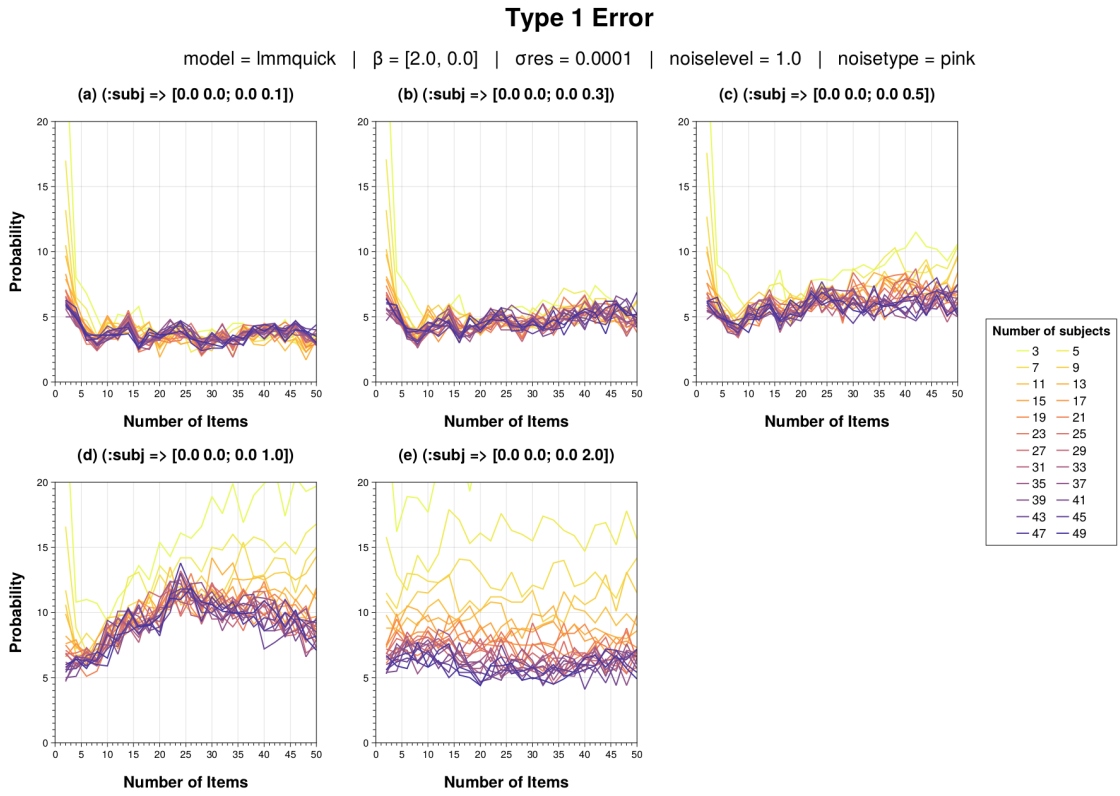


Figure 6.6: Type I error of the LMM (Z-test) with random subject slope variance

6.2 Power Analysis - Two-stage versus LMM

This section describes the observed results of the conducted power analysis for the two-stage approach and the linear mixed models approach. The results are separated into four different subsections. We first take a look at the statistical power for simulated data without random effects. The following two sections look separately at the effects of a random subject intercept or a random subject slope to the statistical power of each modelling scheme. The fourth section combines both, and looks at the results for simulated data with a random subject intercept and a random subject slope. For each section power contour plots for the two-stage approach (left) and the linear mixed model approach (middle) as well as a difference plot (right) are illustrated. The power contour plots below (first & second from the left) display the statistical power for the different number of subjects and items. The number of subjects is shown on the x-axis. The y-axis visualizes the number of items. The wider contour line shows the desired level of 80% statistical power. The difference plot to the right shows the difference between both approaches computed as power of the LMM minus power of the two-stage approach. Red thus shows an advantage of the linear mixed model approach and blue an advantage of the two-stage approach.

6.2.1 No between-subject variation

Figure 6.7 shows the statistical power for the effect size $\beta = 0.5$ and no random effects. For both modelling schemes, it is visible that the statistical power increases with a growing number of subjects and a growing number of items. In the two-stage approach, the statistical power is above 80% for five subjects with 35 items each. The linear mixed model reaches the admired 80% for five subjects with 30 items. For a higher number of subjects, this difference diminishes. The difference plot underlines this observation. For subject numbers three and five, the LMM outscores the two-stage approach with 5 to 10%.

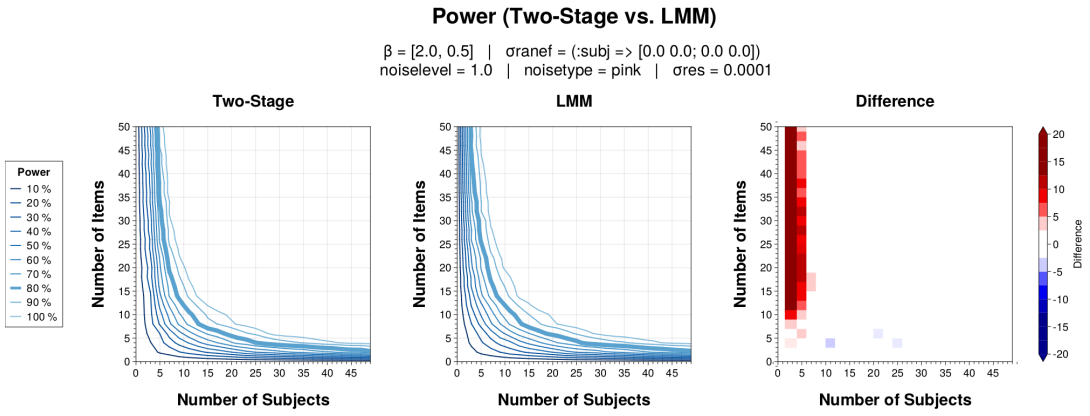


Figure 6.7: Power of two-stage vs LMM approach with no between subject variation

Figure 6.8 visualizes the statistical power for smaller effect sizes ($\beta = 0.3, 0.2, 0.1$). The effect size has, as expected, a large effect on the statistical power. The smaller the effect size, the more items and subjects are needed to reach a statistical power of 80%. This is reflected in the power contour plot by translating the contour lines to the right and top — furthermore, the needed number of subjects to reach the next power level increases.

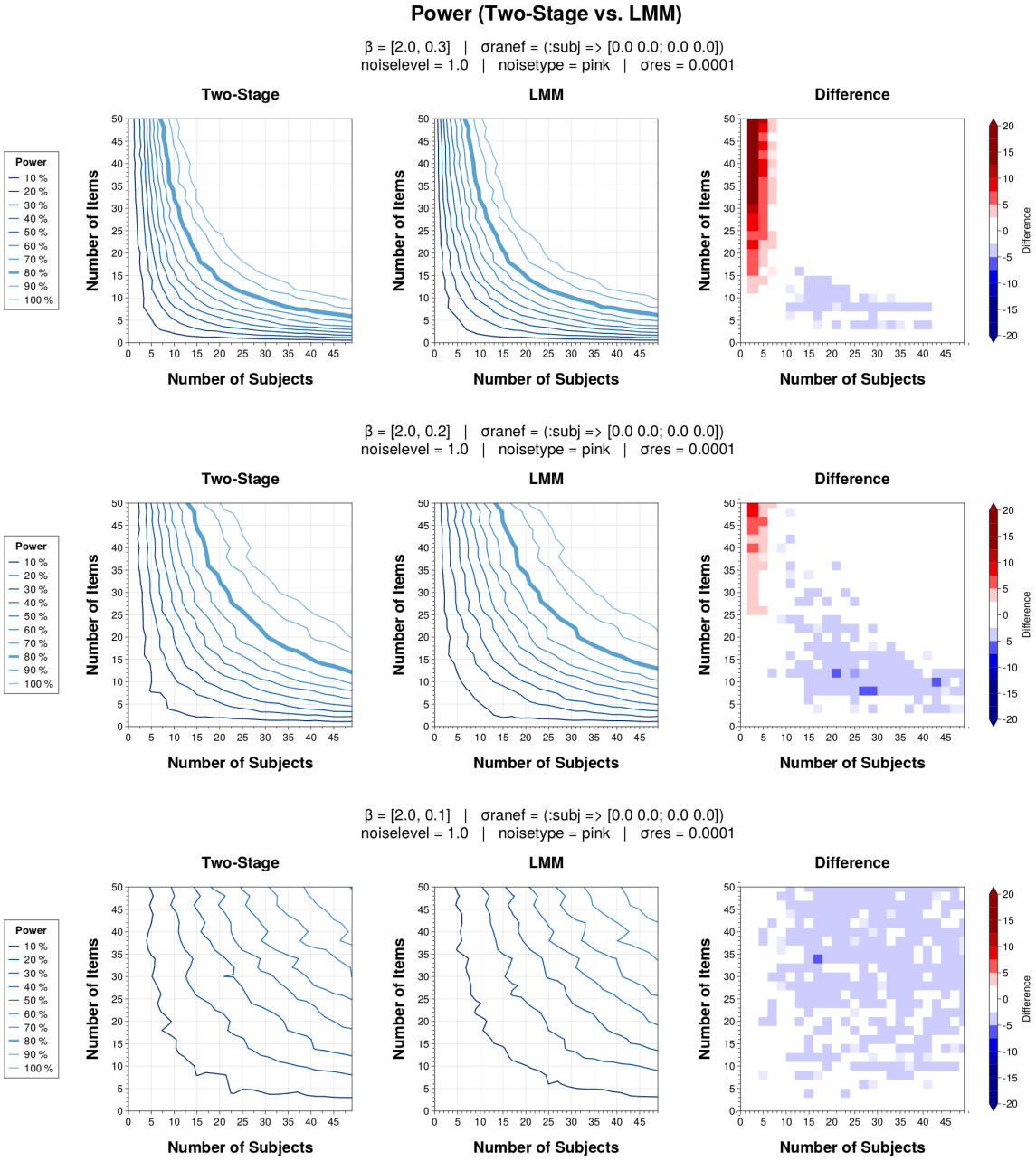
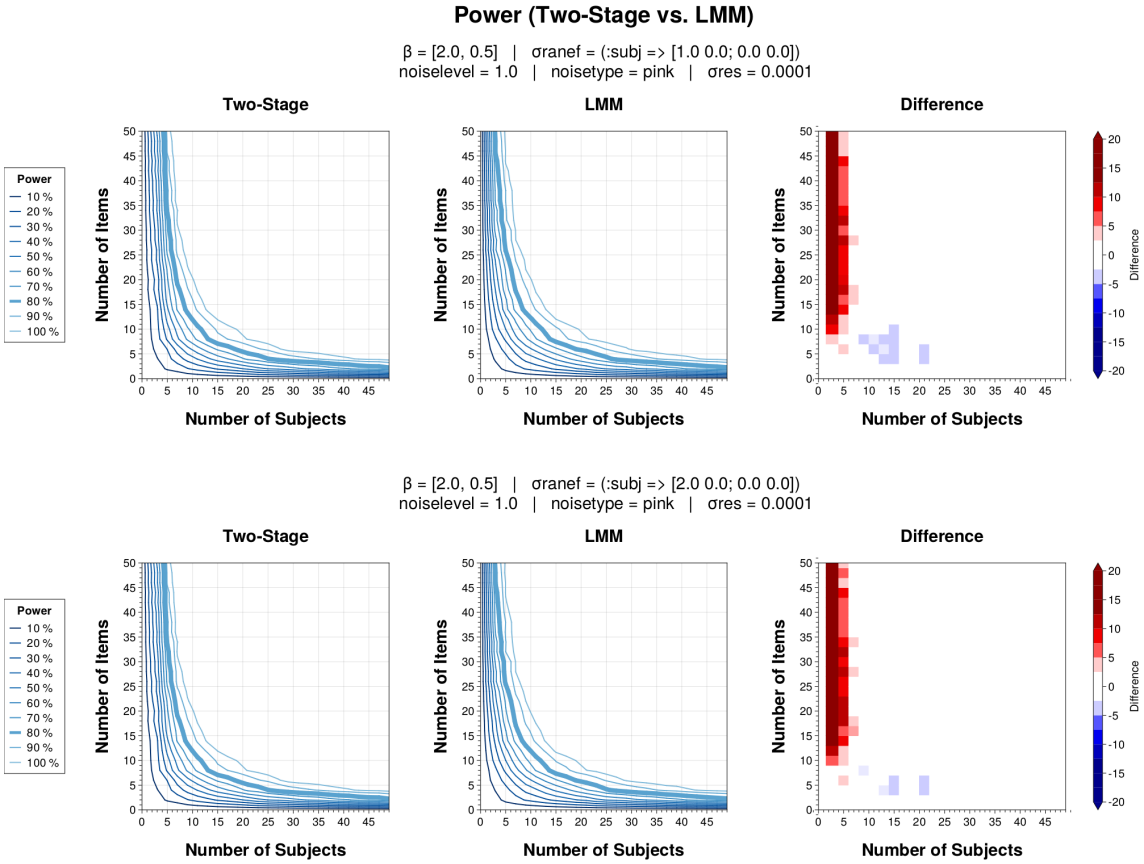


Figure 6.8: Power of two-stage vs LMM approach with no between subject variation for different effect sizes

For an effect size above 0.1, the desired 80% power can be reached with a combination of the number of subjects and items below 50. For an effect size of 0.1, this is with no random effects already impossible for the plotted parameter configuration. Overall, each difference plot shows some differences between both modelling schemes, but the observed differences are negligibly small. This is also refelected in the similarity of the power contour plots.

6.2.2 Random subject intercept

In the following, we will look at the effect of varying between-subject variance in the form of a random intercept per subject. In Figure 6.9, the effect size β is set to 0.5. From the first to third row, the variance of the random subject slope increases from 1.0 over 2.0 to 3.0. The increasing subject intercept variance does not affect the statistical power. Despite some random variations below 5%, the difference plots show the same characteristics as without any random effects. The LMM again performs better for three and five subjects. For a number of subjects greater than five this effect diminishes.



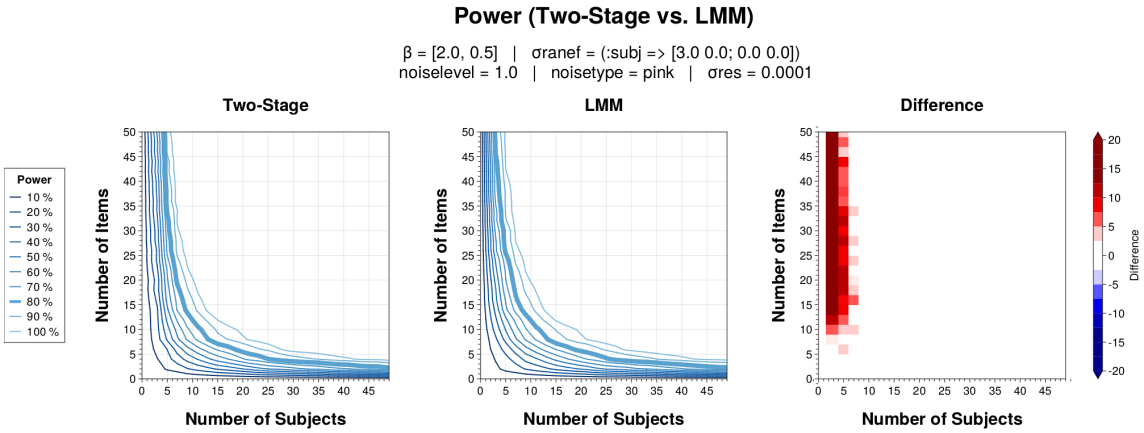


Figure 6.9: Power of the two-stage vs LMM approach with random subject intercept variance

Figure 6.10 shows the power as a function of the number of items for an experiment design with three subjects. Different random intercept slope variances (1.0, 2.0 and 3.0) are visualized by different colors. For a number of three subjects, the LMM performs significantly better. With an increasing number of items, the LMM approach reaches higher statistical power than the two-stage approach. The two-stage approach reaches for a number of items equal to fifty, around 50% statistical power. In contrast, the linear mixed model approach reaches for a number of three subjects and fifty items statistical power above 80%. There is no effect of the random subject variance recognizable .

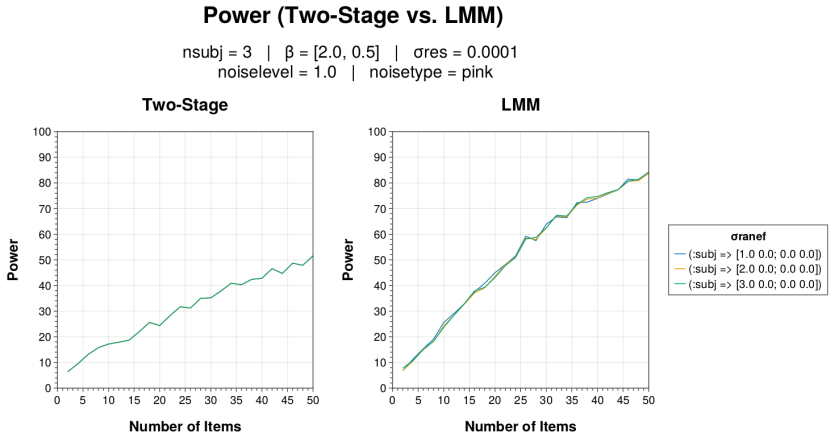


Figure 6.10: Power plot for indivdual number of subjects (three) with random intercept variance

The individual power plot for an experiment design with five subjects, confirms the observation that already with only two subjects more the observed advantage of the LMM reduces and nearly diminishes. Figure 6.11 still shows that the LMM reaches a desired statistical power of 80% faster, but in contrast to with only three subjects, the two-stage approach reaches also a statistical power above 80%.

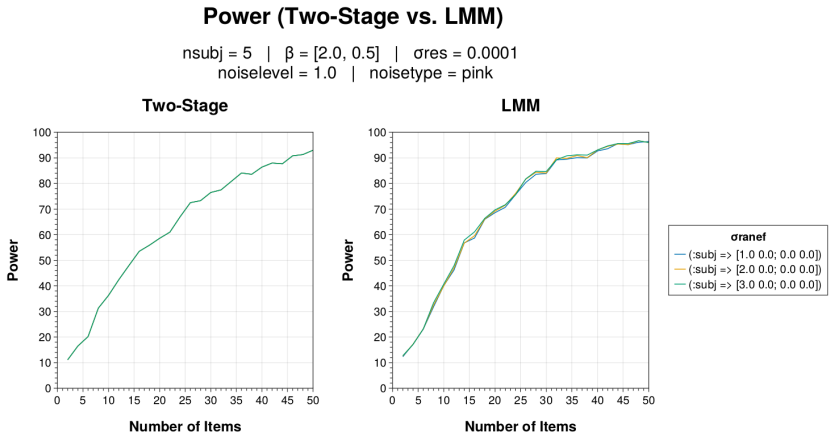


Figure 6.11: Power plot for indivdual number of subjects (five) with random intercept variance

6.2.3 Random subject slope

The next section takes a look at the effect of the random subject slope variance on the power of both modelling schemes. Figure 6.12 shows the power contour plots for the random subject slope variance of 0.5, 1.0 and 2.0. The variance increases from top to bottom. By comparing the power contour plots, it becomes visible that the increase in the random slope variance per subject leads to a translation of the contour lines to the right. Hence, the higher the variance, the more subjects are needed to achieve power above 80%. For a variance of 0.5 (top row), 80% power is reached with a minimum of seven subjects. In the case of the variance equal to 1.0 (second row from top), the power is higher than 80% only for experiment design with more than eleven subjects. A variance of 2.0 (bottom row) requires already more than 35 subjects. The effect of the number of items on the statistical power decreases with higher subject slope variance. For a subject slope variance of 2.0, an experiment design with more than 20 items brings no significant improvement.

The difference plots between the modelling schemes are similar to the difference plots of no random effects and only subject intercept variance. It is no clear improvement for the LMM visible, despite the higher power for subjects three and five. For the variance of 1.0 (middle), the difference between the two-stage approach and LMM is bigger compared to the other variances. With a random subject slope variance of 1.0, the LMM achieves higher statistical power also for experiment designs with seven and nine subjects. With a variance of 2.0 (bottom row), the difference between the modelling schemes shrinks again.

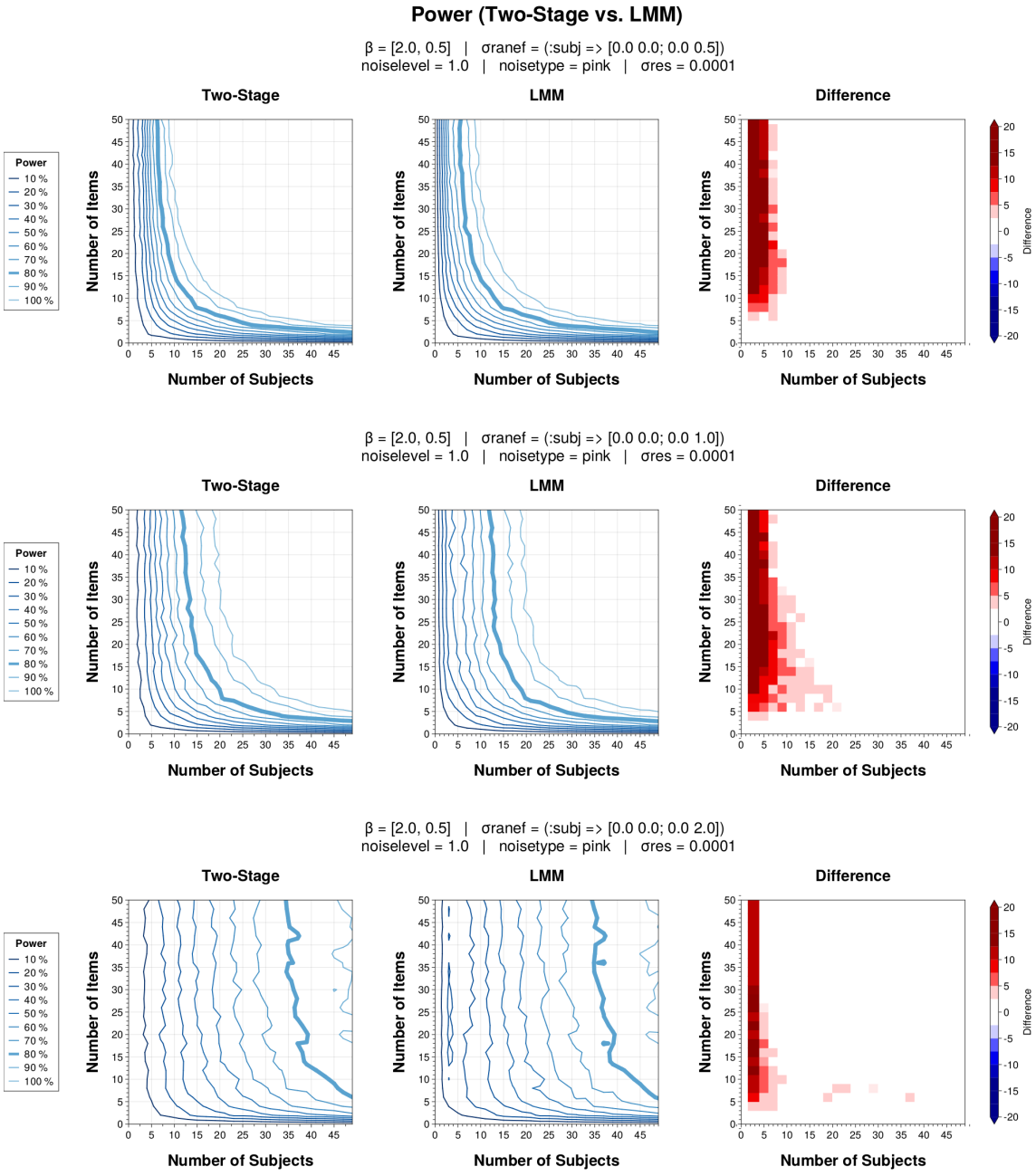


Figure 6.12: Power of the two-stage vs LMM approach with random subject slope variance

The individual power plots for experiment designs with three and five subjects (Figures 6.13 and 6.14 show the power as a function of the number of items for different random subject slope variances (0.1, 0.3, 0.5, 1.0 and 2.0). For a small number of subject and increased subject slope variance both approaches struggle to reach power above 80%. For three subjects, the LMM performs better (Figure 6.13). Despite not reaching the admired level of 80% power, the LMM reaches higher power than the two-stage approach for all variances.

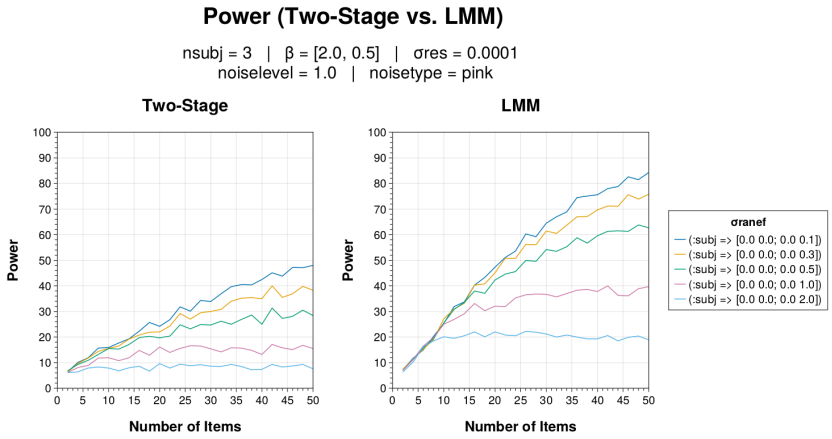


Figure 6.13: Power plot of an experiment design with three subjects for different random slope variances

For an experiment design with five subjects this advantage of the LMM diminishes (Figure 6.14). The two-stage approach reaches for the variances 0.1, 0.3, 0.5, 1.0 and 2.0 similar power as the linear mixed model approach. For only small random subject slope variance (0.1 and 0.3) both approaches reach power above 80% with an number of subjects below 50.

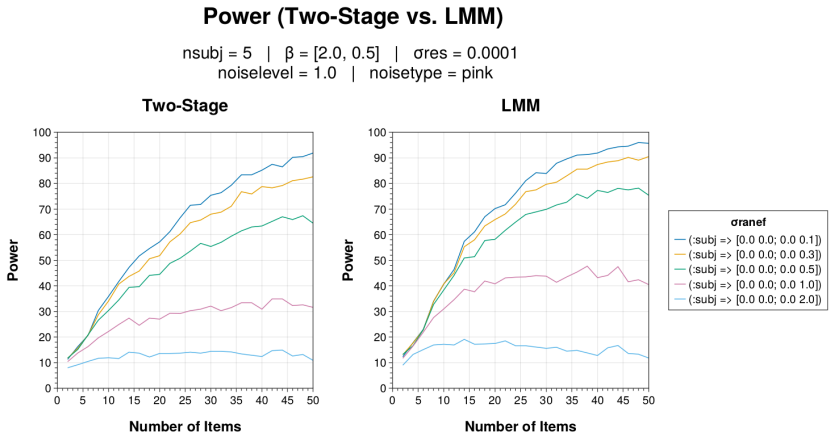


Figure 6.14: Power plot of an experiment design with five subjects for different random slope variances

6.2.4 Random subject intercept and slope

The introduction of a random subject intercept and a random subject slope in the simulated data reveals a similar picture as with only a random slope per subject. The variance of the subject intercept again has little to no effect on the statistical power in both modelling schemes (Figure 6.15).

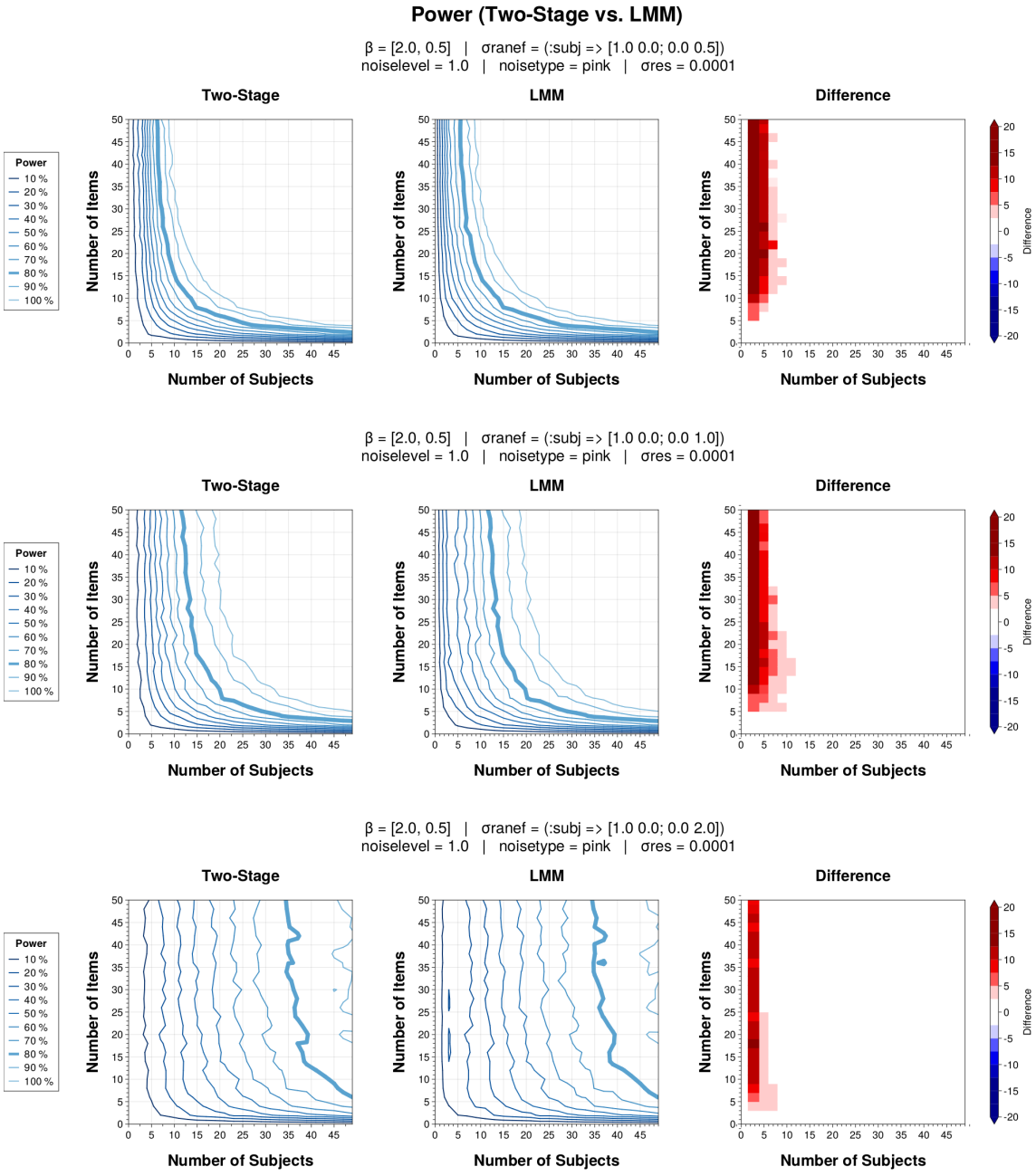


Figure 6.15: Power of two-stage vs LMM approach with random subject intercept and slope

7 Discussion

The thesis aimed to create a simulation toolbox for event-related potentials and conduct a power analysis to compare the two-stage and the linear mixed model approach on simulated data. In the following section, we take a closer look at the results and their limitations and give an outlook on possible further work. We will first focus on the UnfoldSim.jl toolbox and proceed with the conducted comparison.

The created toolbox UnfoldSim.jl provides an intuitive approach to simulate EEG data with subject and item effects. UnfoldSim.jl simplifies the setup for further power analyses and comparisons between different modelling schemes and will hopefully help to improve the power of conducted studies in the field of computational cognitive science. Nonetheless, the toolbox also has some limitations in its current state. Currently, only balanced designs and non-overlapping ERP components can be simulated. Additionally, the simulation is limited to simulating only a single EEG channel. The goal is to steadily improve the scope of functions covered by the toolbox in the future. Optional future goals are implementing an unbalanced design, varying latency for different ERPs, and overlapping ERPs.

The conducted comparison of the two-stage approach and the linear mixed model approach was separated into four experiments – simulations with no random effects, simulations with varying subject intercepts, simulations with varying subject slope and simulations with varying subject intercepts and slopes. The type I error and the statistical power were computed for each experiment.

The observed type I error for the two-stage approach was around the expected level of 5% for data with varying random subject intercepts as well as random subject slopes. For the linear mixed model approach, the same does not apply. Although for data with random subject intercept, the type I error rate was smaller than 5%, the type I error increased above 5% with introducing a random subject slope. Especially for smaller experiment designs with three and five subjects, the type I error was far above 5% the higher the number of items was. For a subject slope variance of 1.0, the observed type I error was above 5% for all subjects. With an increased variance of 2.0, this effect diminished. The type I error of the Z-test for LMMs, showed the same effect, but the increase was more severe. The cause for this increase is unclear and should be further investigated. The simulation and analysis pipeline should thereby also be examined again to exclude errors.

The observed results for the power analysis generally reflect the expected effects of different effect sizes and between-subject variance on the statistical power reported (Baker et al., 2021). The smaller the effect size, the more subjects and items are needed to reach a higher power. The variation in the subject intercept has little to no effect on the statistical power of the analysis. The variation of the random slope per subject leads to more needed subjects for the statistical power to be above 80%.

Nevertheless, the results of the conducted power analysis on simulated EEG data did not present a significant advantage of using linear mixed models instead of the two-stage approach. We expected the linear mixed model approach to outperform the two-stage approach in cases where only a small number of subjects or items are used. Neither for different considered subject intercept variances nor different subject slope variances this was the case. For no random effects, a small advantage (>7.5%) of the linear mixed model approach could be found for experiment designs with three and five subjects. The difference was also visible for simulated data with a varying random subject intercept, random subject slope variance and simulated data with varying subject intercept and slope. This increased effect is likely caused by the high type I error for experiment designs with a small number of subjects. For a random subject slope variance of 1.0, LMMs also showed slightly higher power for experiment design with seven and nine subjects. With a random subject slope variance of 2.0, the increased difference between the modelling schemes diminishes again. This supports the assumption that the observed difference between the linear mixed model and the two-stage approach for a variance of 1.0 is caused by the high type I error since for a variance of 2.0, the effect of an increased type I error decreases. Therefore, the results should be considered with caution.

Besides the increased type I error, other aspects must be considered when interpreting the power analysis results. Unlike the power analyses that were conducted by Baker et al. (2021), the comparison solely relies on simulated data. This introduces the problem of transferring the results to the practice and poses the question of how applicable the results are to real ERP analyses. Further research is needed there. Furthermore, the experiments cover only a small set of parameters due to time limitations and computational costs. The used experiment design and basis function are very rudimentary. The comparison includes only a handful of different effect sizes and between-subject variances. Additionally, the number of subjects and items in the experiments ranges only between 0 and 50. A broader parameter space must be inspected to draw a more founded conclusion. Additionally, the number of subjects for the following analyses should be increased to investigate smaller effect sizes and higher between-subject variances. The same applies to the number of items.

Despite the mentioned limitations, UnfoldSim.jl builds a foundation for various possible future analyses. Besides expanding the current exploratory experiments regarding the between-subject variability, further analyses can focus on comparing models concerning an unbalanced design,

the within-subject variability, the signal-to-noise ratio, and item effects. Especially unbalanced designs and high within-subject variability could be circumstances where the linear mixed model has advantages against the two-stage approach in the ERP analysis.

8 Conclusion

The low statistical power of many recently conducted EEG/ERPs studies is concerning. With experiment designs expanding increasingly fast to more complex situations, commonly used methods should be reviewed and compared to new methods. To our knowledge, a comparison of the two-stage approach and the linear mixed model approach in ERP analysis in a simulated scenario has not been thoroughly explored. The toolbox UnfoldSim.jl forms a framework and a good starting point to change that. The thesis concentrated on the number of subjects and items and the between-subject variability concerning the above-mentioned modelling schemes. Based on the observed results, LMMs did not show a significant advantage for different number of subjects and items and different between-subject variances. However, this should not be a reason to undervalue the use of LMMs in analysing event-related potentials. The comparison was only of exploratory nature and based on a small set of parameters, leaving open space for further research.

Bibliography

Alday, P., D. Bates, L. DeBruine, P. José Bayoán Santiago Calderón, J. Choe, L. Schwetlick (Dec. 2022). *RePsychLing/MixedModelsSim.jl: v0.2.7*. DOI: 10.5281/ZENODO.7407741. URL: https://zenodo.org/record/7407741 (visited on 12/13/2022) (cit. on p. 20).

Alday, P., B. Ehinger, J. Frossard (Mar. 2022). *palday/MixedModelsPermutations.jl: v0.1.4*. DOI: 10.5281/ZENODO.6336248. URL: https://zenodo.org/record/6336248 (visited on 12/13/2022) (cit. on p. 22).

Baayen, R., D. Davidson, D. Bates (Nov. 2008). "Mixed-effects modeling with crossed random effects for subjects and items". en. In: *Journal of Memory and Language* 59.4, pp. 390–412. ISSN: 0749596X. DOI: 10.1016/j.jml.2007.12.005. URL: https://linkinghub.elsevier.com/retrieve/pii/S0749596X07001398 (visited on 05/09/2022) (cit. on p. 15).

Baker, D. H., G. Vilidaite, F. A. Lygo, A. K. Smith, T. R. Flack, A. D. Gouws, T. J. Andrews (June 2021). "Power contours: Optimising sample size and precision in experimental psychology and human neuroscience." en. In: *Psychological Methods* 26.3, pp. 295–314. ISSN: 1939-1463, 1082-989X. DOI: 10.1037/met0000337. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/met0000337 (visited on 05/11/2022) (cit. on pp. 4, 38).

Barr, D. J., R. Levy, C. Scheepers, H. J. Tily (Apr. 2013). "Random effects structure for confirmatory hypothesis testing: Keep it maximal". en. In: *Journal of Memory and Language* 68.3, pp. 255–278. ISSN: 0749596X. DOI: 10.1016/j.jml.2012.11.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S0749596X12001180 (visited on 12/10/2022) (cit. on pp. 5, 16, 17, 22).

Bates, D., P. Alday, D. Kleinschmidt, P. José Bayoán Santiago Calderón, L. Zhan, A. Noack, M. Bouchet-Valat, A. Arslan, T. Kelman, A. Baldassari, B. Ehinger, D. Karrasch, E. Saba, J. Quinn, M. Hatherly, M. Piibeleht, P. K. Mogensen, S. Babayan, Y. L. Gagnon (Oct. 2022). *JuliaStats/MixedModels.jl: v4.7.3*. DOI: 10.5281/ZENODO.7153199. URL: https://zenodo.org/record/7153199 (visited on 12/13/2022) (cit. on pp. 20, 22).

Boudewyn, M. A., S. J. Luck, J. L. Farrens, E. S. Kappenman (June 2018). "How many trials does it take to get a significant ERP effect? It depends". en. In: *Psychophysiology* 55.6, e13049. ISSN: 00485772. DOI: 10.1111/psyp.13049. URL: https://onlinelibrary.wiley.com/doi/10.1111/psyp.13049 (visited on 12/10/2022) (cit. on p. 4).

Button, K. S., J. P. A. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. J. Robinson, M. R. Munafò (May 2013). "Power failure: why small sample size undermines the reliability of neuroscience". en. In: *Nature Reviews Neuroscience* 14.5, pp. 365–376. ISSN: 1471-003X, 1471-0048. DOI: 10.1038/nrn3475. URL: http://www.nature.com/articles/nrn3475 (visited on 06/03/2022) (cit. on p. 4).

Daniel, W., C. Cross (2013). *Biostatistics: A Foundation for Analysis in the Health Sciences*. Wiley Series in Probability and Statistics. Wiley. ISBN: 978-1-118-30279-8. URL: https://books.google.de/books?id=5I5jMAEACAAJ (cit. on pp. 8, 9).

Ehinger, B. V., O. Dimigen (2019). "Unfold: An integrated toolbox for overlap correction, non-linear modeling, and regression-based EEG analysis". In: *peerJ*. DOI: 10.7717/peerj.7838 (cit. on p. 22).

Heise, M. J., S. K. Mon, L. C. Bowman (Apr. 2022). "Utility of linear mixed effects models for event-related potential research with infants and children". en. In: *Developmental Cognitive Neuroscience* 54, p. 101070. ISSN: 18789293. DOI: 10.1016/j.dcn.2022.101070. URL: https://linkinghub.elsevier.com/retrieve/pii/S1878929322000147 (visited on 05/09/2022) (cit. on p. 5).

HypothesisTests.jl (2022). URL: https://github.com/JuliaStats/HypothesisTests.jl (visited on 12/13/2022) (cit. on p. 22).

Jackson, A. F., D. J. Bolger (Nov. 2014). "The neurophysiological bases of EEG and EEG measurement: A review for the rest of us: Neurophysiological bases of EEG". en. In: *Psychophysiology* 51.11, pp. 1061–1071. ISSN: 00485772. DOI: 10.1111/psyp.12283. URL: https://onlinelibrary.wiley.com/doi/10.1111/psyp.12283 (visited on 12/13/2022) (cit. on p. 6).

Kappenman, E. S., J. L. Farrens, W. Zhang, A. X. Stewart, S. J. Luck (Jan. 2021). "ERP CORE: An open resource for human event-related potential research". en. In: *NeuroImage* 225, p. 117465. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2020.117465. URL: https://linkinghub.elsevier.com/retrieve/pii/S1053811920309502 (visited on 12/13/2022) (cit. on p. 7).

Laird, N. M., J. H. Ware (1982). "Random-Effects Models for Longitudinal Data". In: *Biometrics* 38.4. Publisher: [Wiley, International Biometric Society], pp. 963–974. ISSN: 0006-341X. DOI: 10.2307/2529876. URL: https://www.jstor.org/stable/2529876 (visited on 12/10/2022) (cit. on p. 15).

Luck, S. (2014). *An Introduction to the Event-Related Potential Technique, second edition*. A Bradford Book. MIT Press. ISBN: 978-0-262-52585-5. URL: https://books.google.de/books?id=SzavAwAAQBAJ (cit. on pp. 6, 7).

Matuschek, H., R. Kliegl, S. Vasishth, H. Baayen, D. Bates (June 2017). "Balancing Type I error and power in linear mixed models". en. In: *Journal of Memory and Language* 94, pp. 305–315. ISSN: 0749596X. DOI: 10.1016/j.jml.2017.01.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S0749596X17300013 (visited on 12/10/2022) (cit. on pp. 5, 17, 22, 23).

OpenScienceCollaboration (2015). “Estimating the reproducibility of psychological science”. In: *Science* 349.6251. _eprint: https://www.science.org/doi/pdf/10.1126/science.aac4716, aac4716. DOI: 10.1126/science.aac4716. URL: https://www.science.org/doi/abs/10.1126/science.aac4716 (cit. on p. 4).

Pavlov, Y. G., N. Adamian, S. Appelhoff, M. Arvaneh, C. S. Benwell, C. Beste, A. R. Bland, D. E. Bradford, F. Bublatzky, N. A. Busch, P. E. Clayson, D. Cruse, A. Czeszumski, A. Dreber, G. Dumas, B. Ehinger, G. Ganis, X. He, J. A. Hinojosa, C. Huber-Huber, M. Inzlicht, B. N. Jack, M. Johannesson, R. Jones, E. Kalenkovich, L. Kaltwasser, H. Karimi-Rouzbahani, A. Keil, P. König, L. Kouara, L. Kulke, C. D. Ladouceur, N. Langer, H. R. Liesefeld, D. Luque, A. MacNamara, L. Mudrik, M. Muthuraman, L. B. Neal, G. Nilsonne, G. Niso, S. Ocklenburg, R. Oostenveld, C. R. Pernet, G. Pourtois, M. Ruzzoli, S. M. Sass, A. Schaefer, M. Senderecka, J. S. Snyder, C. K. Tamnes, E. Tognoli, M. K. van Vugt, E. Verona, R. Vloeberghs, D. Welke, J. R. Wessel, I. Zakharov, F. Mushtaq (Nov. 2021). “#EEGManyLabs: Investigating the replicability of influential EEG experiments”. en. In: *Cortex* 144, pp. 213–229. ISSN: 00109452. DOI: 10.1016/j.cortex.2021.03.013. URL: https://linkinghub.elsevier.com/retrieve/pii/S0010945221001106 (visited on 06/03/2022) (cit. on p. 4).

Simmons, J. P., L. D. Nelson, U. Simonsohn (Nov. 2011). “False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant”. en. In: *Psychological Science* 22.11, pp. 1359–1366. ISSN: 0956-7976, 1467-9280. DOI: 10.1177/0956797611417632. URL: http://journals.sagepub.com/doi/10.1177/0956797611417632 (visited on 06/03/2022) (cit. on p. 4).

Smith, N. J., M. Kutas (2015). “Regression-based estimation of ERP waveforms: I. The rERP framework”. In: *Psychophysiology* 52.2, pp. 157–168. ISSN: 14698986. DOI: 10.1111/psyp.12317 (cit. on p. 14).

Warrington, N. M., K. Tilling, L. D. Howe, L. Paternoster, C. E. Pennell, Y. Y. Wu, L. Briollais (Jan. 2014). “Robustness of the linear mixed effects model to error distribution assumptions and the consequences for genome-wide association studies”. In: *Statistical Applications in Genetics and Molecular Biology* 13.5. ISSN: 1544-6115, 2194-6302. DOI: 10.1515/sagmb-2013-0066. URL: https://www.degruyter.com/document/doi/10.1515/sagmb-2013-0066/html (visited on 12/12/2022) (cit. on p. 17).

Wilkinson, G. N., C. E. Rogers (1973). “Symbolic Description of Factorial Models for Analysis of Variance”. In: *Applied Statistics* 22.3, p. 392. ISSN: 00359254. DOI: 10.2307/2346786. URL: https://www.jstor.org/stable/10.2307/2346786?origin=crossref (visited on 12/14/2022) (cit. on p. 12).

Erklärung

Ich versichere, diese Arbeit selbstständig verfasst zu haben. Ich habe keine anderen als die angegebenen Quellen benutzt und alle wörtlich oder sinngemäß aus anderen Werken übernommene Aussagen als solche gekennzeichnet. Weder diese Arbeit noch wesentliche Teile daraus waren bisher Gegenstand eines anderen Prüfungsverfahrens. Ich habe diese Arbeit bisher weder teilweise noch vollständig veröffentlicht. Das elektronische Exemplar stimmt mit allen eingereichten Druck-Exemplaren überein.

Datum und Unterschrift:

Declaration

I hereby declare that the work presented in this thesis is entirely my own. I did not use any other sources and references that the listed ones. I have marked all direct or indirect statements from other sources contained therein as quotations. Neither this work nor significant parts of it were part of another examination procedure. I have not published this work in whole or in part before. The electronic copy is consistent with all submitted hard copies.

Date and Signature: