

A Tutorial on Tailored Simulation-Based Sample Size Planning for Experimental Designs  
with Generalized Linear Mixed Models

Timo K. Koch<sup>†,1, 2</sup>, Florian Pargent<sup>†,1</sup>, Anne-Kathrin Kleine<sup>1</sup>, Eva Lermer<sup>1,3</sup>, & Susanne  
Gaube<sup>1,4</sup>

<sup>1</sup> LMU Munich, Department of Psychology

<sup>2</sup> University of St. Gallen, Institute of Behavioral Science & Technology

<sup>3</sup> Technical University of Applied Sciences Augsburg, Department of Business Psychology

<sup>4</sup> University College London, Global Business School for Health

## Author Note

This research was funded by a grant from the Volkswagen Foundation (Grant No. 98525). All materials (reproducible manuscript incl. R code, simulation results) are available in the project's repository on the Open Science Framework (OSF):

<https://osf.io/dhwf4/>

The authors made the following contributions. Timo K. Koch: Conceptualization, Formal Analysis, Writing - Original Draft Preparation, Writing - Review & Editing; Florian Pargent: Formal Analysis, Methodology, Visualization, Writing - Original Draft Preparation, Writing - Review & Editing; Anne-Kathrin Kleine: Formal Analysis, Writing - Review & Editing; Eva Lerner: Conceptualization, Resources, Writing - Review & Editing; Susanne Gaube: Conceptualization, Supervision, Writing - Review & Editing. <sup>†</sup> Timo K. Koch and Florian Pargent contributed equally to this work.

Correspondence concerning this article should be addressed to Timo K. Koch, Torstrasse 25, 9000 St. Gallen, Switzerland. E-mail: [timo.koch@unisg.ch](mailto:timo.koch@unisg.ch)

### Abstract

When planning experimental research, determining an appropriate sample size and using suitable statistical models are crucial for robust and informative results. However, the recent replication crisis underlines the need for more rigorous statistical methodology and well-powered designs. Generalized linear mixed models (GLMMs) offer a flexible statistical framework to analyze experimental data with complex (e.g., dependent and hierarchical) data structures. Yet, available methods and software for a priori power analyses for GLMMs are often limited to specific designs, while data simulation approaches offer more flexibility. Based on a practical case study, the current tutorial equips researchers with a step-by-step guide and corresponding code for conducting tailored a priori power analyses to determine appropriate sample sizes with GLMMs. Finally, we give an outlook on the increasing importance of simulation-based power analysis in experimental research.

*Keywords:* power analysis, data simulation, sample size, generalized linear mixed model

Word count: 7504

A Tutorial on Tailored Simulation-Based Sample Size Planning for Experimental Designs  
with Generalized Linear Mixed Models

## Introduction

When planning experimental research, it is essential to determine an appropriate sample size to ensure that the results obtained are both robust and informative, and to use appropriate statistical models to analyze the data (Lakens, 2022b). However, the recent replication crisis has illustrated many challenges surrounding the reproducibility and reliability of findings (Yarkoni, 2022). As a result, there is a growing need for more rigorous statistical methodology and the adoption of well-powered experimental designs. While easy-to-use software solutions exist for simple statistical models and experimental designs, many researchers lack the skills and tools to conduct “a priori” (i.e., before data collection) sample size planning for more complex research designs using the flexible generalized linear mixed models (GLMM) framework in order to determine the required sample size in their experiments. In the present work, we provide a tutorial on performing tailored simulation based sample size planning for GLMMs. After introducing general concepts, we review existing software solutions in R and discuss under which circumstances tailored data simulations are necessary. Then we describe general steps and decisions that are necessary whenever performing tailored data simulation. Because the details of these steps vary considerably for each specific study, we finish with a case study from the field of human-AI (artificial intelligence) interaction research.

—>

## Prior Knowledge

TODO: fertig ausformulieren

Simulation-based sample size planning for generalized linear models is a complex topic that touches on a variety of statistical concepts and programming skills. Although we give

some short introduction to clarify our terminology, to profit the most from our tutorial, our readers require prior knowledge in the following topics (or refer to our collection of background literature):

Readers should understand basic statistical concepts like hypothesis tests and their statistical power but also confidence intervals and their precision. Some background knowledge in empirical estimands, statistical estimates, and causal inference is very useful but not absolutely necessary.

In addition, readers should have basic programming skills in R on how to simulate data and how to repeat these simulations in parallel to speed up computation. While there are a myriad of options for data simulation in R beyond the basic functionality in the base and stats packages, in our tutorial we use functions from the tidyverse and the faux package.

Finally, readers should be familiar with regression modeling in general and GLMMs in particular. In our tutorial, we simulate data by manually specifying the model equation of a GLMM that represents our assumed data generating process. While it is not necessary to understand the technical details of how GLMMs are estimated, it is crucial to understand the structure of a basic GLMM (e.g., logistic regression with random intercepts) and how the model assumes that the values in the dependent variable are determined by the predictor variables and the random effects.

#### Literature:

- Lakens ebook (Lakens, 2022a) evtl. Andy Field für die Basics?
- R for data science (Wickham, Çetinkaya-Rundel, & Golemund, 2023), Bruine article (DeBruine & Barr, 2021), faux documentation (DeBruine, 2023) für die R skills
- kumle power tutorial (Kumle, Vö, & Draschkow, 2021) und irgendwelche paper (am besten aus AMPPS) zu GLMMs
- what is your estimand? paper (Lundberg, Johnson, & Stewart, 2021) für das estimand

topic

## General concepts

### Planning for statistical power or precision

Conducting research with too small sample sizes can have many negative consequences. First, experiments may yield inconclusive or misleading results, hindering the accumulation of knowledge. Second, studies that are doomed to never finding a postulated effect due to small sample size waste resources by consuming time, effort, and funding without delivering meaningful results. For such reasons, many journals and funding agencies now require that a justification of sample size is included in study protocols and grant proposals, recognizing its significance in ensuring robust and meaningful findings. Although many pragmatic justifications are possible (Lakens, 2022b), the ideal is often considered to determine a suitable sample size a priori (i.e., before the study is conducted) based on some meaningful computation to assure that the study will be able to fulfill its purpose. In this way, adding a solid sample size calculation to the research process can act as a safeguard for ensuring high-quality research.

—>

The majority of empirical studies in psychology applies hypothesis testing. As a consequence, the dominant approach for sample size planning is based on power analysis (i.e., planning for power). However, sample size planning can also be based on the precision of parameter estimates (i.e., planning for precision).

**Planning for power.** In empirical research relying on hypothesis testing, the most common strategy for determining an adequate sample size is based on statistical power (Lakens, 2022b). Statistical power is defined as the probability that a hypothesis test has a significant p-value when analyzing repeated samples from a population with a true effect of some pre-specified size. Less formally, power is described as the probability that a hypothesis

test correctly rejects the null hypothesis when the alternative hypothesis is true. If the sample size (i.e., the number of participants and/or stimuli) used for data collection is insufficient to detect the effects or relationships being investigated with high probability, the study would be considered “underpowered”.

When planning for power, a target is set for the statistical power of a hypothesis test of interest (a very common heuristic is 0.8). Assuming some effect size of interest and a desired significance level, a minimum sample size can be determined that (on average) would guarantee reaching this target.

**Planning for precision.** Not all research questions are best answered using hypothesis testing. Especially for exploratory studies when little previous research has been conducted, only estimating effects of interest is often more useful. When no hypothesis tests are conducted when analysing the planned study, power analysis is not available for sample size planning. Nonetheless, sample size still has a crucial effect on how informative the planned study will be to learn about the effect of interest. In this case, the expected width of a confidence interval is used as a target for sample size planning. The values inside a confidence interval are commonly interpreted as plausible values for the quantity of interest the confidence interval is supposed to estimate. More formally, a 95%-confidence interval provides the smallest interval with the property that upon repeated sampling, 95% of individual confidence intervals would include the true quantity of interest. Thus, a narrow confidence interval (assuming the confidence level is kept at a high value) is more informative about the size of the true effect than a wide interval. Apart from the desired confidence level, the width of a confidence interval depends strongly on the sample size. Because bigger samples carry more information, they lead to smaller confidence intervals (assuming everything else stays the same). When planning for precision, a target is set for the expected width of a confidence interval of interest. Assuming some effect size of interest and a desired confidence level, a minimum sample size can be determined that (on average) would guarantee reaching this target.

## Generalized linear mixed models (GLMMs)

As study designs become more complex, psychological researchers require more sophisticated statistical models to capture the nuanced relationships and grouping structures introduced by their study designs (Yarkoni, 2022). GLMMs (also called multilevel models) are gaining increasing popularity in analyzing data because they offer a flexible framework for handling data with outcome variables that are not normally distributed (e.g., categorical outcomes) while accounting for both fixed and random effects (Fahrmeir, Kneib, Lang, & Marx, 2021).

GLMMs are an extension of LMMs (Linear Mixed Models), which are, in turn, extensions of linear regression models that account for correlated data including hierarchical structures (Fahrmeir et al., 2021). In this context, correlated data means that the value in the outcome variable for one observation may be related (i.e., more similar or less similar) to the value of another observation in a systematic way that is not already accounted for by the usual (fixed) predictor variables (e.g., age of participants). This correlation can arise for various reasons: Responses to some stimuli from some participants might be more similar because the same person was measured twice (repeated measurements), both participants come from the same neighborhood (clustering) or both participants responded to the same stimulus (stimulus effects). Thus, modeling such correlations is especially important whenever the data has a clear structure, while the grouping variables can be hierarchically organized (e.g., students nested in schools, schools nested in districts) or not (e.g., students solve math exercises, but neither student sees all exercises). LMMs are used when the outcome variable is continuous and follows a normal distribution (when conditioned on all fixed and random effects). They allow for the modeling of fixed effects, which capture the relationships between our usual predictors and the outcome, as well as random effects, which account for the different types of correlation structure and grouping effects exemplified above. Random effects are typically assumed to follow a normal distribution with a mean of zero



and a variance that quantifies the heterogeneity across groups.

As mentioned, GLMMs extend the LMM framework to accommodate non-normally distributed continuous and categorical outcome variables. GLMMs incorporate both fixed and random effects, similar to LMMs, but also involve a link function that connects the linear combination of predictor variables to the expected value of the outcome variable. The link function allows for modeling the relationship between predictors and the outcome in a non-linear way that is appropriate for the specific distribution family of the outcome variable. As an example, think of an experiment with different design factors (e.g., picture, headline) impacting the likelihood of users clicking on an online advertisement. Here, participants' behavior is measured repeatedly (e.g., over several sessions). The click patterns of participants in one session are likely to be correlated with their previous sessions. Finally, the outcome variable is binary (click/no click) for each session, which follows a binomial distribution.

### **Simulation based sample size planning with GLMMs**

For simpler statistical models, like t-tests, ANOVA, and linear regression, with common study designs (e.g., mean comparison between two groups), user-friendly software for a priori sample size planning is readily available (Champely et al., 2018; Faul, Erdfelder, Buchner, & Lang, 2009; Kelley, Maxwell, & Rausch, 2003; Kelley & Rausch, 2006). However, these software packages are often not flexible enough to perform sample size planning for complex designs.

Existing approaches for sample size planning with GLMMs have almost exclusively focused on planning for power. Power analysis methods for multilevel models can be categorized into formula-based methods and simulation-based methods (Murayama, Usami, & Sakaki, 2022). Formula-based methods rely on often complicated formulas that can be used to directly calculate power while simulation-based methods rely on repeatedly

simulating data with a known true effect size and estimating power empirically (i.e., how often the hypothesis test is significant for the simulated data). Currently available formula-based software packages for power analysis with multilevel models often do not include GLMMs or are limited to very simple designs (Murayama et al., 2022; Westfall, Kenny, & Judd, 2014), making it necessary to build data simulations tailored specifically to the study design. A number of tutorials have been published describing how to perform such simulation-based power analysis for multilevel models (Arend & Schäfer, 2019; Brysbaert & Stevens, 2018; DeBruine & Barr, 2021; Kumle et al., 2021; Lafit et al., 2021; Zimmer, Henninger, & Debelak, 2022). However, most of these tutorials focus on linear mixed models (LMMs) and the most common designs (but see Kumle et al., 2021 for a tutorial that also covers more advanced settings). This narrow focus provides limited guidance for researchers faced with more complex study designs, especially when little prior knowledge about plausible effect sizes is available (see the discussion in Kumle et al., 2021). The necessary presumptions for simulation-based power analysis with GLMMs include assumptions about the distributional form of the outcome variable, the random effects, and the correlation structure. The distributional assumption specifies the distributional family for the outcome variable (when conditioned on all fixed and random effects). Assumptions about the random effects include the assumption of normality (i.e., that the random effects follow a normal distribution) and the covariance structure among the random effects (i.e., if and how they are correlated). Interpreting these presumptions entails understanding the underlying presumptions of the model and ensuring they align with the characteristics of the data being analyzed. Existing tutorials often rely on heuristics for specifying variance components (e.g., the standard deviation of random intercepts) or assume that results from meta-analyses or data from pilot studies are available to determine plausible values for all model parameters. However, in practice, knowledge about those parameters from prior studies is often limited, which makes specifying assumptions a practical challenge Kumle et al. (2021).

- Insert table on existing R packages for power analysis with GLMMs

### **When do we profit from tailored data simulation?**

Performing tailored simulation based sample size planning is more complicated and time consuming than using existing software tools for sample size planning. Thus, it is worth discussing under which circumstances we most profit from or are forced to perform tailored simulation based sample size planning, because existing software solutions do not meet our requirements.

### **Complex study designs**

TODO: fertig ausformulieren

Real studies are often more complex than the simplified designs assumed by many user-friendly software packages for sample size planning.

**Missing data.** TODO: fertig ausformulieren

Missing data is a frequent issue in applied data analysis. There can be various reasons for why data is missing. For example data can be missing completely at random. Alternatively, subjects might drop out or produce missing data as a result of some covariate also measured in the dataset. As a different note, many experimental designs contain some conditions in which some values of the predictor variables are missing by design. Missing by design conditions can make data analysis more complicated because predictors have to be coded in highly specific ways that prevent the estimated GLMM to become unidentified. Whether missing data has an effect on the result of sample size planning depends on the characteristics of the missing data process. However, except for very simple scenarios, deciding whether missing data can be safely ignored in the sample size planning is often difficult to decide on a theoretical basis. A better strategy is to explicitly include the (assumed) process of missing data into the data simulation process.

**Data generating processes beyond standard GLMMs.** TODO: fertig ausformulieren

While GLMMs are a very flexible model class that can handle a large variety of outcome variables, psychological researchers are becoming increasingly aware that many psychological datasets might profit from even more sophisticated models. Common examples are zero-inflated outcomes, censoring, and nonlinear predictor effects.

- possibility to extend beyond standard GLMMs (e.g. zero-inflated outcomes, censoring, n

### Complex statistical hypotheses

TODO: fertig ausformulieren

The most common hypotheses tested in psychological research are of the type  $H_0 : \beta = 0$ , where  $\beta$  is some coefficient in a regression model. However, many research questions in psychology actually require testing more complex statistical hypotheses.

**Directed hypotheses and testing against values other than 0.** TODO: fertig ausformulieren

In the new era of preregistration and registered reports, most research questions are tested with directed hypotheses, because good theories should at least postulate whether some psychological effect of interest is positive or negative. Even better theories should be able to specify a smallest effect sizes of interest (SESOI) that must be exceeded if the psychological effect has some practical relevance in a specific application. In combination, this might require a test like  $H_0 : \beta \leq 0.1$ .

**Combined hypotheses.** TODO: fertig ausformulieren

Elaborate research questions often require testing hypotheses that consist of a combination of model parameters, for example  $H_0 : \beta_0 + \beta_1 \leq 0$ . If the research question consists only of a single hypothesis of this sort, it is often possible to reduce the hypothesis to a single regression coefficient by clever coding and/or centering of predictor variables. However, interesting research questions often consist of combined hypotheses that consist of

more than one separate statistical hypotheses. For example, a combined null hypothesis  $H_0$  might consist of two single null hypotheses  $H_{01} : \beta_1 \leq 0$  and  $H_{02} : \beta_0 + \beta_1 \leq 0$ . For some research questions, the combined null hypothesis  $H_0$  would be rejected if both  $H_{01}$  AND  $H_{02}$  are rejected. For other research questions, the combined null hypothesis  $H_0$  would be rejected if  $H_{01}$  OR  $H_{02}$  OR both are rejected. If the global hypothesis  $H_0$  is combined with OR, the p-values of the single hypotheses must be corrected for multiple testing to avoid  $\alpha$ -inflation for the global hypothesis. However, if the global hypothesis  $H_0$  is combined with AND, a correction for multiple testing is not necessary but rather a mistake that unnecessarily reduces the power of the global hypothesis test.

None of the software packages for sample size planning in table X can handle combined hypotheses as discussed here. In contrast, our case study will demonstrate how we can test directed combined hypotheses with tailored simulation based sample size planning.

### Planning for precision

TODO: fertig ausformulieren

As discussed earlier, the two major frameworks for sample size planning are planning for power and planning for precision. Although planning for precision has been increasingly discussed as a more suitable strategy for exploratory research, all available software packages for sample size planning with GLMMs are based on power analysis.

As long as one were interested only in confidence intervals of a single model parameter, it might be possible to adapt existing software packages to plan for precision. However, we demonstrate in our case study that in realistic research projects with GLMMs, the contrasts of interests and their respective confidence intervals are usually situated on the scale of the (expected) outcome in contrast to the scale of the model parameters. We will show that tailored simulation based sample size planning can easily handle the planning for precision scenario. The only change in procedure is that instead of computing hypothesis tests for

each simulated dataset and estimating statistical power across repetitions, confidence intervals are computed for each simulated dataset and the expected width is estimated.

### **No prior studies or pilot data**

TODO: fertig ausformulieren

All frameworks for sample size planning require the user to make assumptions about the expected effect size. Assuming the true effect is of this size (or greater), one can compute the (minimum) power of a hypothesis test or the (maximum) expected width of a confidence interval. Existing software packages for sample size planning for GLMMs usually require to provide the assumed effect in the unit of some standardized measure of effect size. When the researcher has access to similar studies or pilot data, providing such standardized effect sizes is feasible. However, this requirement can be an almost impossible challenge when no prior studies or pilot data are available. This problem is further exacerbated by the fact that GLMMs are so flexible, that general heuristics of what should be considered a small effect do not exist or are difficult to defend. In our experience, using domain knowledge to construct a tailored data simulation is the only solution to determine plausible effect sizes in the absence of prior evidence. It would be possible to use these tailored simulations to extract plausible values for standardized effect sizes that could then be inserted in existing software packages for sample size planning. However, we would argue that when tailored data simulations are necessary anyways, performing the whole desing analysis in a customized way is preferred over using the existing software packages.

### **General steps in tailored simulation-based sample size planning**

Although the details differ depending on the specific study characteristics, each tailored simulation based sample size planning requires a series of steps and decisions that we will outline in the remainder of our tutorial. Will introduce each step in a theoretical section, followed by the practical application based on a case study. All code in this

manuscript and simulation results are available in the project’s repository on the Open Science Framework (<https://osf.io/dhwf4/>).

## Define the estimand

**THEORY.** The first step in every research process should always be a clear definition of which theoretical quantity is necessary to answer the specific research question of the study under planning. Unfortunately, many psychological researchers are still not well trained in this task. The theoretical estimand is a unit specific quantity defined outside of any statistical model (Lundberg et al., 2021). Depending on the research and whether the research question is causal, this quantity might be counterfactual and not directly observable. Part of the theoretical estimand is also a clear definition of the target population, for which the quantity is of interest.

When the theoretical estimand has been specified, the next step is to define the empirical estimand, which reformulate the theoretical estimand in a way that can be estimated based on the observed data that will be collected in the planned study. For many common research questions in psychology, the empirical estimand can be expressed as a statistical quantities within the scope that can be estimated with a regression model. Note however, that this is not always possible and that the literature on causal inference includes many estimation strategies beyond regression (CITATION).

When the empirical estimand has been specified, the last step is to clearly define how a concrete estimate of the empirical estimand shall be computed based on the observed study data. When a regression model is used for estimation, this includes decisions like which statistical framework estimation technique, and software package is used to fit the model, and how to compute hypothesis tests or confidence intervals.

**PRACTICE.** In the present case study, we consider the effectiveness of feedback provided by an artificial intelligence (AI) in an AI-enabled diagnostic decision support

system. The context is a clinical setting, where expert radiologists and students under training must detect bleeding based on head scans from computer tomography. In the investigated AI-enabled diagnostic decision support system, an AI model can provide initial diagnostic advice, which can be used as guidance by the humans that are required to make the final diagnostic decision. The research goal is to validate the effectiveness of the AI-enabled advice. We consider the AI-enabled advice as effective, if the following pattern holds, which we will first describe verbally:

*We expect that for BOTH expert radiologists and students under training, correct AI-advice leads to a higher probability of accurately diagnosing a CT scan compared to no AI-advice presented, AND, we expect that for BOTH experts and non-experts, incorrect advice leads to a lower probability of accurately diagnosing a CT scan compared to no advice presented.*

It becomes clear that our theoretical estimand is be a causal quantity, based on several individual quantities that are defined outside of any statistical model (Lundberg et al., 2021). To elaborate further, we focus on only one part of the research question concerning the comparison between correct AI-advice and no AI-advice for expert radiologists. For a specific expert radiologist and a specific CT scan, our question is concerned with the difference between the probability that a correct diagnosis is made if correct AI-advice were presented and the probability that a correct diagnosis is made if no AI-advice were presented. Mathematically, we could describe this unit-specific quantity with some pseudo-equation like this (where  $s$  stands for a specific expert radiologist and  $i$  stands for a specific CT scan):

$$P(\text{correct diagnosis}|\text{correct advice, specific expert, specific scan}) \\ - P(\text{correct diagnosis}|\text{no advice, specific expert, specific scan})$$

However, our verbal description is not precise enough yet, because the above quantity may differ for each combination of expert and scan. We must specify how we want to aggregate



the quantity across different experts and scans. If we would compute the average of our individual quantity across all experts and scans in the target population, the result would be called the average treatment effect (ATE) in the population of experts (CITATION).

Although often debated whether this is a good idea, psychological research often focuses on a different quantity: Instead of averaging over experts and scans, we focus in a hypothetical *typical* expert and a *typical* scan, where *typical* is usually defined as an average score on all attributes of the expert or scan. With this definition, the theoretical estimand in our above example can be expressed as:

$$\begin{aligned} &P(\text{correct diagnosis}|\text{correct advice, average expert, average scan}) \\ &\quad - P(\text{correct diagnosis}|\text{no advice, average expert, average scan}) \end{aligned}$$

In total, our theoretical estimand consists of four such expressions, the remaining three are:

$$\begin{aligned} &P(\text{correct diagnosis}|\text{no advice, average expert, average scan}) \\ &\quad - P(\text{correct diagnosis}|\text{wrong advice, average expert, average scan}) \end{aligned}$$

$$\begin{aligned} &P(\text{correct diagnosis}|\text{correct advice, average student, average scan}) \\ &\quad - P(\text{correct diagnosis}|\text{no advice, average student, average scan}) \end{aligned}$$

$$\begin{aligned} &P(\text{correct diagnosis}|\text{no advice, average student, average scan}) \\ &\quad - P(\text{correct diagnosis}|\text{wrong advice, average student, average scan}) \end{aligned}$$

To complete our definition of the theoretical estimand, we have to clearly define our target population that consists of both persons and scans: With respect to persons, we are only interested in expert radiologists and medical students at German universities. With respect to stimuli, we are only interested in the head CT scans made from subjects that do or do not suffer from intracerebral hemorrhage. Similarly, we are only interested in AI-advice given by the specific AI-enabled diagnostic decision support system under study.

While a theoretical estimand is defined outside of any statistical model and can be impossible to be observe directly, an empirical estimand is a quantity that can be estimated based on observed data. For our exemplary research question, it is possible to construct an experimental study, where all participants are confronted with the same set of brain scans, but the kind of AI-advice given for each brain scan is randomly assigned within participants. This random intervention allows to estimate our theoretical estimand, although in reality, each person receives only one kind of AI-advice (correct advice, wrong advice, no advice) for each brain scan. For our example, each probability expression in our theoretical estimand can be modeled with the same GLMM. Estimating the GLMM based on the data observed in our planned study produces an estimate for each probability and these estimates can be used to compute an estimate for each of the four probability contrasts. For pedagogical reasons, we will skip the concrete definition of our empirical estimand until we have discussed how to simulate data based on a concrete GLMM in the next section.

### **Simulate the data generating process**

***THEORY.*** When the estimand has been defined, the next step in the research process is to write code that simulates the (assumed) data generating process of the planned study. This requires specifying a generative process for all predictor variables used in the final data analysis. While such assumptions can be quite challenging for observational studies and/or continuous predictor variables, this is less of a problem for experimental studies with only categorical predictor variables. When all predictor variables have been simulated, one can use the structure of a suitable GLMM to simulate the dependent variable. To simulate the GLMM, one requires plausible values for all model parameters. We will discuss strategies on how these values can be obtained later. Because one has full control over the data generating process in a tailored simulation, it is simple to also model highly specific aspects of the planned study, like data missing by design or assuming that subjects drop out (i.e. deleting their data from the simulated dataset) based on other variables in the

simulated dataset. The quality of the results of the sample size planning crucially depends on the plausibility of the simulated data generating process. For this reason, deciding on a specific set of assumptions can easily feel overwhelming. However, we would argue that even a strongly simplified data generating process (e.g. including only a small number of interaction effects; including only random intercepts but no random slopes; making rather implausible assumptions for missing data) can yield informative result and should be strongly preferred to the current status quo in psychological research, where most studies do not justify their sample size or use highly questionable heuristics that are independent of their specific research question.

***PRACTICE.*** In our case study, we simulate data for an experiment where the diagnostic performance of users of an AI-enabled diagnostic decision support system will be evaluated. Participants, radiologists (task experts) and students/interns (non experts), review a series of head computer tomography (CT) scans to assess the presence of a bleeding. To support their decision-making, an AI model provides initial diagnostic advice, which can be used as guidance by the participants. In the control condition, no AI advice is presented, meaning that the participants have to read the CT scan without any support. When AI advice is given, this advice can be either correct or incorrect. The type of advice (no advice, wrong advice, correct advice) is randomized within subjects across brain scans. After reviewing a CT scan, participants deliver a medical diagnosis (bleeding or no bleeding), which may be either accurate or inaccurate. This experimental design introduces some missing values by design since the advice is neither correct nor incorrect when no advice is present, which must be taken into account when simulating and analyzing the data. In this example, recruiting task experts (i.e., radiologists) is more challenging due to their limited availability, while non-experts (i.e., students/interns) are more readily accessible. The goal of simulation-based sample size planning is to determine how many task experts and non-experts must be recruited to achieve sufficient statistical power or precision in the planned experiment.

*Our specific GLMM.* In a GLMM, the expected value of the dependent variable  $Y$  conditioned on the vector of predictor variables  $\mathbf{X}$  and random effects  $\mathbf{U}$ , transformed by a link function  $g(\cdot)$  is modeled as a linear combination  $\eta$  of the predictor variables  $\mathbf{X}$ , the random effects  $\mathbf{U}$  and the model parameters  $\beta$  (Fahrmeir et al., 2021):

$$g(E(Y|\mathbf{X} = \mathbf{x}, \mathbf{U} = \mathbf{u})) = \eta$$

Equivalently, the conditional expected value is modeled as the linear combination  $\eta$ , transformed by the inverse link function  $g^{-1}(\cdot)$ :

$$E(Y|\mathbf{X} = \mathbf{x}, \mathbf{U} = \mathbf{u}) = g^{-1}(\eta)$$

If the dependent variable (i.e., diagnostic decision)  $Y$  is a binary variable with values 0 (i.e., inaccurate), or 1 (i.e., accurate), the conditional expected value is equivalent to the probability:

$$P_{si} := P(Y = 1|\mathbf{X} = \mathbf{x}, \mathbf{U} = \mathbf{u})$$

In our case study,  $P_{si}$  is the conditional probability that subject  $s$  gives the correct response to item (i.e., CT scan)  $i$ .

In such a setting, we model this probability as

$$P_{si} = \text{inverse\_logit}(\eta_{si})$$

with the inverse-logit link  $g^{-1}(\eta_{si}) = \text{inverse\_logit}(\eta_{si}) = \frac{\exp(\eta_{si})}{1+\exp(\eta_{si})}$  or equivalently

$$\text{logit}(P_{si}) = \eta_{si}$$

with the logit link  $g(P_{si}) = \text{logit}(P_{si}) = \ln(\frac{P_{si}}{1-P_{si}})$ .

In our case study, the probability of making an accurate diagnostic decision is assumed to depend on the predictors:

- *advice\_present<sub>si</sub>*: whether subject  $s$  was presented with AI advice (1) or not (0) when asked to assess item  $i$

- $advice\_correct_{si}$ : whether this advice was correct (1) or not (0)
- $expert_s$ : whether subject  $s$  was a task expert (1) or not (0)

and the random effects:

- $u_{0s}$ : the deviation of subject  $s$  from the average ability to solve an item (i.e., CT scan) with average difficulty; assumed to be distributed as  $u_{0s} \sim N(0, \sigma_S^2)$
- $u_{0i}$ : the deviation of item (i.e., CT scan)  $i$  from the average difficulty to be solved by a person with average ability; assumed to be distributed as  $u_{0i} \sim N(0, \sigma_I^2)$

In total, we assume the model

$$\begin{aligned} \text{logit}[P_{si}] = & (\beta_0 + u_{0s} + u_{0i}) + \\ & \beta_a \cdot advice\_present_{si} + \beta_c \cdot advice\_correct_{si} + \beta_e \cdot expert_s + \\ & \beta_{ea} \cdot expert_s \cdot advice\_present_{si} + \beta_{ec} \cdot expert_s \cdot advice\_correct_{si} \end{aligned}$$

or equivalently

$$\begin{aligned} P_{si} = \text{inverse\_logit}[ & (\beta_0 + u_{0s} + u_{0i}) + \\ & \beta_a \cdot advice\_present_{si} + \beta_c \cdot advice\_correct_{si} + \beta_e \cdot expert_s + \\ & \beta_{ea} \cdot expert_s \cdot advice\_present_{si} + \beta_{ec} \cdot expert_s \cdot advice\_correct_{si} ] \end{aligned}$$

with model parameters  $\beta_0$ ,  $\beta_e$ ,  $\beta_a$ ,  $\beta_c$ ,  $\beta_{ea}$ ,  $\beta_{ec}$ ,  $\sigma_S$ , and  $\sigma_I$ .

In the GLMM literature, this would be called a binomial GLMM with two random intercepts (for subjects and items), two level-1 predictors ( $advice\_present$ ,  $advice\_correct$ ), one level-2 predictor ( $expert$ ) and two cross-level interactions ( $expert \cdot advice\_present$ ,  $expert \cdot advice\_correct$ ). To limit complexity, we do not consider random slopes, additional predictors or higher-level interactions.

*Simulation function in R.* The following R function simulates a full dataset structured according to the design of our case study. The `faux` package (DeBruine, 2023) contains useful functions when simulating factorial designs, including random effects.

```

simulate <- function(n_subjects = 100, n_items = 50,
  b_0 = 0.847, b_e = 1.350, b_a = -1.253, b_c = 2.603,
  b_ea = 0.790, b_ec = -1.393,
  sd_u0s = 0.5, sd_u0i = 0.5, ...){
  require(dplyr)
  require(faux)

  # simulate design
  dat <- add_random(subject = n_subjects, item = n_items) %>%
    add_between("subject", expert = c(1, 0), .prob = c(0.25, 0.75)) %>%
    mutate(advice_present = rbinom(n(), 1, prob = 2/3)) %>%
    mutate(advice_correct = if_else(advice_present == 1L,
                                   rbinom(n(), 1L, prob = 0.8), 0L)) %>%

  # add random effects
  add_ranef("subject", u0s = sd_u0s) %>%
  add_ranef("item", u0i = sd_u0i) %>%

  # compute dependent variable
  mutate(linpred = b_0 + u0i + u0s +
    b_e * expert + b_a * advice_present + b_c * advice_correct +
    b_ea * expert * advice_present + b_ec * expert * advice_correct) %>%
  mutate(y_prob = plogis(linpred)) %>%
  mutate(y_bin = rbinom(n = n(), size = 1, prob = y_prob))

  dat
}

```

In the first six lines of the function definition, we set some default parameter values (which we will explain in the next section) and load the packages we use to manipulate and simulate data. In our case study, each subject (`n_subjects` in total) is assumed to respond

to each item (i.e., CT scan; `n_items` in total). Thus, the `add_random` command creates a fully-crossed `data.frame` with `n_subjects`  $\times$  `n_items` rows. We add a between-subject effect with the `add_between` command, simulating that about 25% of subjects are experts. The next two lines simulate that in  $\frac{2}{3}$  of trials, subjects will be presented with AI advice, and if advice is presented, the advice will be correct in about 80% of cases (the variable `advice_correct` is always 0 when no advice is presented). Next, we simulate one random effect for each subject (`u0s`) and for each item (`u0i`). As assumed by standard GLMMs, the `add_ranef` function draws the random effects from a normal distribution with a mean 0 and a standard deviation specified by the user. With all design variables done, we are ready to simulate our model equation outlined in the last section. The linear predictor variable `linpred` ( $\eta$  in the GLMM model equations) combines the predictor variables, random effects, and model parameters as assumed by our model. We then transform the linear predictor with the inverse-link function to compute `y_prob`, the probability that the subject correctly solved the item (in R, the inverse-logit link is computed with `plogis` and the logit link with `qlogis`). In the final step, we simulate the binary dependent variable `y_bin` (i.e., whether the subject makes an accurate diagnostic decision for the CT scan) by – for each trial – drawing from a Bernoulli distribution with success probability `y_prob`.

### Strategies on specifying the population parameters

**THEORY.** Tailored simulation based sample size planning is especially relevant when studies have to be planned in the absence of previous studies with the same design or pilot data. In this scenario, researchers require strategies on how to specify the population parameters used in their data simulation. Population parameters are all model parameters estimated in a GLMM, in particular the regression coefficients of the fixed effects and the standard deviation of the random effects (and the correlation between random effects in more complicated models). In contrast to non-hierarchical linear regression, common heuristics based on standardized effect sizes are less useful or not even available for GLMMs.

Our strategies to specify population parameters will require access to domain knowledge from domain experts. Because most domain knowledge of domain experts can only be expressed in unstandardized measurement units of a specific application, we argue that unstandardized effect sizes are usually preferable over standardized effect sizes for tailored simulation based sample size planning. The basic idea of all strategies is how the data generating process implied by certain values of population parameters can be quantified or visualized in an intuitive way that enables a calibration of population parameters based on the available knowledge of domain experts.

Although we focus on frequentist statistics in our tutorial, selecting plausible values for data simulation shares many similarities with the specification of (weakly) informative prior distribution in applied Bayesian statistics. In this literature, strategies have been developed to monitor the plausibility of (prior) model assumptions (Gelman et al., 2020), that can be adapted for our purposes.

**PRACTICE.** When introducing the simulation function for our case study, we have used theoretically plausible values as defaults for all model parameters ( $\beta_0$ ,  $\beta_e$ ,  $\beta_a$ ,  $\beta_c$ ,  $\beta_{ea}$ ,  $\beta_{ec}$ ,  $\sigma_S$ , and  $\sigma_I$ ), but have not talked about where these numbers came from.

Ideally, one would rely on meta-analytic results or conclusive data from pilot studies. However, these are sometimes not readily available. All parameter values in our present case study have been determined based on results from distantly related study designs in the literature. Additionally, we had repeated discussions with our affiliated domain experts in radiology to check whether our assumptions seem plausible.

We now outline our main strategy to determine plausible parameter values for the fixed effects ( $\beta$  parameters): Unfortunately, the model parameters in a binomial GLMM are hard to interpret in isolation because 1) the  $\beta$  parameters are connected to the modeled probability via the non-linear inverse-logit link, and 2) we also have to consider the random effects. The most simple interpretation, that allows us to ignore the random effects for now,



works by imagining a subject with average ability ( $u_{0s} = 0$ ) responding to an item (i.e., CT scan) with average difficulty ( $u_{0i} = 0$ ). Then the model implied probability that such a person solves such an item accurately is given by:

$$\begin{aligned}
 P(Y = 1 | \mathbf{X} = \mathbf{x}, \mathbf{U} = \mathbf{0}) = \\
 = \text{inverse\_logit}[\beta_0 + \beta_a \cdot \text{advice\_present}_{si} + \beta_c \cdot \text{advice\_correct}_{si} + \beta_e \cdot \text{expert}_s + \\
 \beta_{ea} \cdot \text{expert}_s \cdot \text{advice\_present}_{si} + \beta_{ec} \cdot \text{expert}_s \cdot \text{advice\_correct}_{si}]
 \end{aligned}$$

In fact, we would only need the full equation if the subject is an expert and correct advice is presented. In all other experimental conditions, some terms drop from the equation because they are multiplied by 0. The other extreme case would be the probability that a non-expert with average ability solves an item with average difficulty when no advice is presented:

$$\begin{aligned}
 P(Y = 1 | \text{advice\_present} = 0, \text{advice\_correct} = 0, \text{expert} = 0, u_{0s} = 0, u_{0i} = 0) = \\
 = \text{inverse\_logit}[\beta_0]
 \end{aligned}$$

We can revert this perspective by choosing plausible probability values based on domain knowledge and deriving the parameter values implied by these probabilities for each experimental condition.

Table 1 shows our set of assumptions concerning the probability that an average subject solves an average item for each experimental condition, as well as the corresponding equations implied by the model. The table can be used to compute the implied values for the  $\beta$  parameters, starting with the first equation and reinserting the computed  $\beta$  values in all following equations:

```

b_0 <- qlogis(0.7)
b_e <- qlogis(0.9) - b_0
b_a <- qlogis(0.4) - b_0
b_ea <- qlogis(0.85) - b_0 - b_e - b_a

```

Table 1

*Assumed probabilities that an average subject solves an average item in each experimental condition.*

Experimental condition	$P(Y = 1 \mathbf{X} = \mathbf{x}, \mathbf{U} = \mathbf{0})$	Implied equation
no advice, no expert	0.70	$\text{logit}(0.70) = \beta_0$
no advice, expert	0.90	$\text{logit}(0.90) = \beta_0 + \beta_e$
false advice, no expert	0.40	$\text{logit}(0.40) = \beta_0 + \beta_a$
false advice, expert	0.85	$\text{logit}(0.85) = \beta_0 + \beta_e + \beta_a + \beta_{ea}$
correct advice, no expert	0.90	$\text{logit}(0.90) = \beta_0 + \beta_a + \beta_c$
correct advice, expert	0.95	$\text{logit}(0.95) = \beta_0 + \beta_e + \beta_a + \beta_c + \beta_{ea} + \beta_{ec}$

*Note.* Implied equations are derived based on the model equations and setting all random intercept terms to 0.

```

b_c <- qlogis(0.9) - b_0 - b_a
b_ec <- qlogis(0.95) - b_0 - b_e - b_a - b_c - b_ea
c(b_0 = b_0, b_e = b_e, b_a = b_a, b_c = b_c, b_ea = b_ea, b_ec = b_ec)

```

```

##          b_0          b_e          b_a          b_c          b_ea          b_ec
## 0.8472979 1.3499267 -1.2527630 2.6026897 0.7901394 -1.3928518

```

It is always possible to double-check these computations by transforming the parameter values back to probabilities, e.g.

$$\begin{aligned}
P(Y = 1 | \text{expert} = 1, \text{advice\_present} = 1, \text{advice\_correct} = 1, u_{0s} = 0, u_{0i} = 0) = \\
= \text{inverse\_logit}[\beta_0 + \beta_e + \beta_a + \beta_c + \beta_{ea} + \beta_{ec}]
\end{aligned}$$

which we compute in R as:

```
plogis(b_0 + b_e + b_a + b_c + b_ea + b_ec)
```

```
## [1] 0.95
```

This leaves us with the question on how to determine plausible values for the two remaining model parameters ( $\sigma_S$ , and  $\sigma_I$ ) that are the standard deviations for the random intercepts. For this, we introduce two more strategies.

### **Insightful descriptive statistics.**

**THEORY.** The mathematical structure of GLMMs determines which patterns in data would be produced by the model, if a specific set of values for the population parameters would be specified. The knowledge of how to simulate from a GLMM enables us to compute insightful descriptive statistics that can be compared to available domain knowledge much more easily than the opaque values of model parameters. For example, domain experts might not be able to directly choose plausible values for the  $\beta$  coefficients in a logistic regression model (which are measured on the log-odds scale). However, they should be able to reason about the expected ratio of the binary dependent variable in different experimental condition, i.e. which descriptive statistics they expect to observe (on average) when they will analyse the actual dataset produced by the study under planning (if everything turns out as postulated by the investigated theory). The job of the analyst who is familiar with the mathematical structure of the GLMM is to produce the model implied value of the insightful descriptive statistic expected by the domain expert. Although insightful descriptive statistics often depend on the model parameters in complicated ways, it is usually possible to manually adjust the population parameters (if only with the help of trial and error) until the model implied quantities produce the desired result.

**PRACTICE.** In the last section, we showed how we can derive the model implied probability for each experimental condition, that a subject with average ability solves an item with average difficulty. Although these derivations are straightforward, it is important not to misinterpret their implications: In binomial GLMMs, the average probability to solve

an item (averaged across persons of varying ability and items of varying difficulty) is **not** equal to the probability that a person with average ability solves an item with average difficulty (Fahrmeir et al., 2021). The first perspective implies a so-called marginal interpretation, while the second one implies a conditional interpretation.

For example, we determined the  $\beta$  parameters in a way that corresponds to a desired conditional probability of 0.95, that an expert with average ability solves an item with average difficulty when presented with correct advice (the conditional perspective). However, even if the model were true, we would not observe that 95% of experts responding to items presented with correct advice from a big sample of subjects drawn from their natural distribution of ability and items drawn from their natural distribution of difficulty (the marginal perspective). How much the two probabilities differ depend on the standard deviations of the random intercepts (the two probabilities are only equal if both standard deviations would be zero). We want to use the model implied observed proportion of correct diagnoses in each experimental condition as an insightful descriptive statistics to determine plausible values for the random effect standard deviations.

We will simulate a large dataset (for which the observed values of the descriptive statistic will be close to their model implied true values) and simply compute the relative frequency of correct diagnoses for each experimental condition.

```
library(tidyverse)
set.seed(1)
dat <- simulate(n_subjects = 3000, n_items = 3000,
  sd_u0s = 0.5, sd_u0i = 0.5)
dat %>%
  mutate(condition = fct_cross(
    factor(expert), factor(advice_present), factor(advice_correct))) %>%
  mutate(condition = fct_recode(condition,
```

```

    "no expert, no advice" = "0:0:0", "expert, no advice" = "1:0:0",
    "no expert, wrong advice" = "0:1:0", "expert, wrong advice" = "1:1:0",
    "no expert, correct advice" = "0:1:1", "expert, correct advice" = "1:1:1")) %>%
group_by(condition) %>%
summarize(relative_frequency = sum(y_bin) / n())

```

```

## # A tibble: 6 x 2
##   condition          relative_frequency
##   <fct>              <dbl>
## 1 no expert, no advice          0.683
## 2 expert, no advice            0.881
## 3 no expert, wrong advice       0.409
## 4 expert, wrong advice          0.828
## 5 no expert, correct advice     0.883
## 6 expert, correct advice        0.938

```

We tried using these descriptive statistics to judge together with our domain experts whether our chosen values for the random effect standard deviations would produce data that align with our domain expertise. However although the result was deemed plausible, in our example these statistics are not informative enough to determine a final set of plausible parameter values. For this reason, we will additionally look at insightful model based quantities.

### Insightful model based quantities.

***THEORY.*** Because GLMMs are complicated models, descriptive statistics alone are usually not enough to specify plausible values for all model parameters. This is especially true for the standard deviation of random effects, that have complicated (and often unexpected) effects on the model-implied results. An important advantage of data simulation (where one has full control over parameter values and sample sizes) is that one can produce

insightful model based quantities that can never be directly observed in an actual empirical dataset. For example, in a logistic model with random intercepts for participants, one can produce a visualization of the implied distribution of the probability that a participant on average solves a cognitive task. Although domain knowledge will probably not suffice to specify this distribution completely, it should be possible to rule out unplausible boundary conditions. For example, the domain expert might deem it unplausible that the 5% most able participants have a probability of more than 0.99 to solve the difficult cognitive task.

***PRACTICE.*** The discussed inequality of conditional and marginal effects in GLMMs (Fahrmeir et al., 2021) makes their interpretation more difficult. One must be careful when specifying parameter values based on previous studies or pilot data that use the marginal interpretation (e.g., a pilot study providing an estimate of how often neurologists make an accurate diagnosis based on brain scans). However, this does not mean that we cannot use the marginal interpretation (average probability across persons and items) to inform plausible parameter values: When parameter values have been selected, we can compute the implied marginal distributions and compare this information to our domain knowledge. Then, we can iteratively adjust the parameter values until we are satisfied with the implied distributions.

In the last section, we simulated a large simulated dataset and computed descriptive statistics, the relative frequencies of correct diagnoses, for each experimental condition. We will now use the model implied probability of each simulated data point (stored in the variable `y_prob`) to visualize the whole model implied marginal distribution of correct diagnoses for each experimental condition.

```
library(ggdist)

dat %>%

  mutate(condition = fct_cross(

    factor(expert), factor(advice_present), factor(advice_correct))) %>%
```

```
mutate(condition = fct_recode(condition,
  "no expert, no advice" = "0:0:0", "expert, no advice" = "1:0:0",
  "no expert, wrong advice" = "0:1:0", "expert, wrong advice" = "1:1:0",
  "no expert, correct advice" = "0:1:1", "expert, correct advice" = "1:1:1")) %>%
ggplot(aes(x = y_prob, y = condition)) +
stat_histinterval(point_interval = "mean_qi", slab_color = "gray45") +
scale_x_continuous(breaks = seq(0, 1, 0.1), limits = c(0, 1))
```

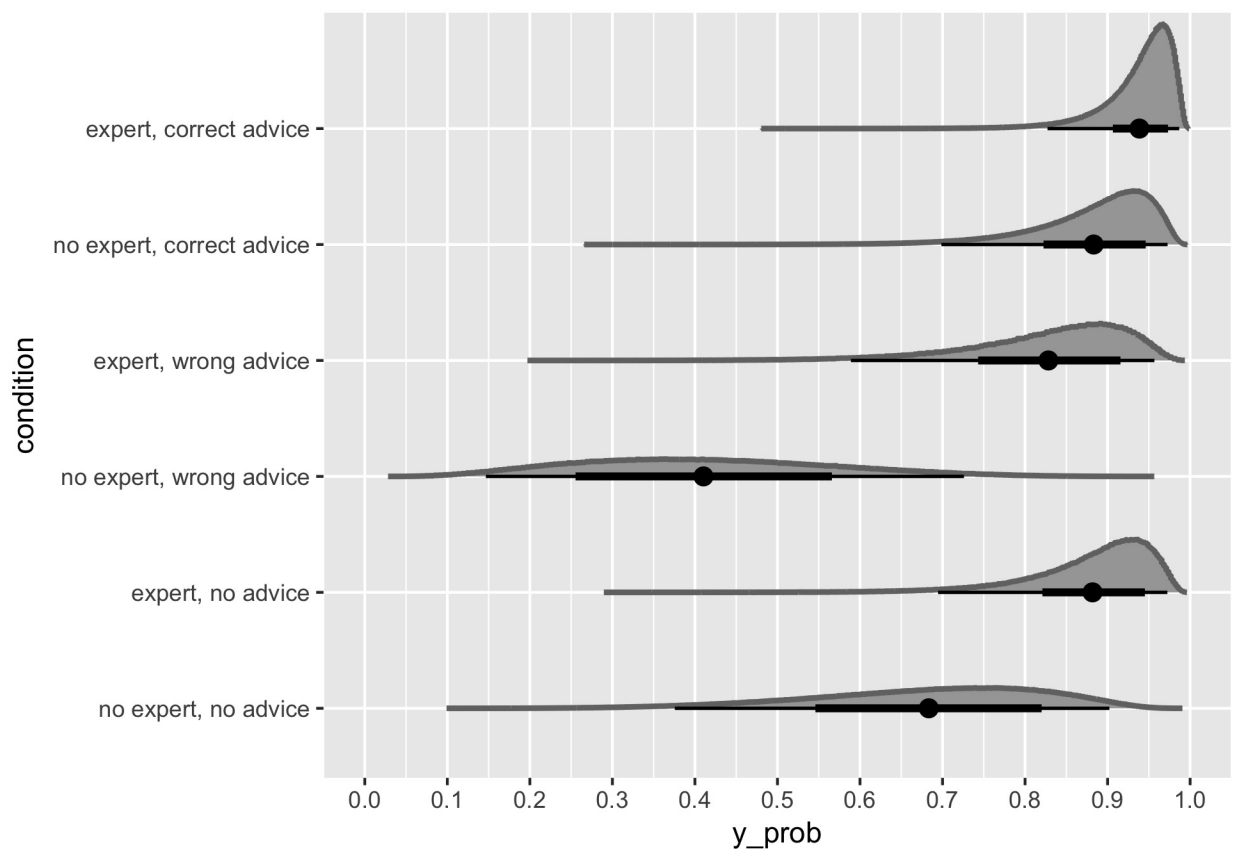


Figure 1. Marginal distributions including means, 66% and 95% confidence intervals for all experimental conditions.

Figure 1 shows the model implied marginal distributions, including the mean, 66% and 95% intervals. We can see that, indeed, the average probabilities (black dots) slightly differ from the probabilities of average subjects and items considered in the previous section. This

difference increases with the variability of the random effects.

We can use plots like the one above as a useful tool to decide whether the specified standard deviations of the subject and item random intercepts ( $\sigma_S$  and  $\sigma_I$ ) are reasonable by comparing the ranges and overlap between conditions to domain knowledge.

In the next plot, we have set the item standard deviation to almost zero ( $\sigma_I = 0.01$ ). This gives us a better way to see the variability between persons.

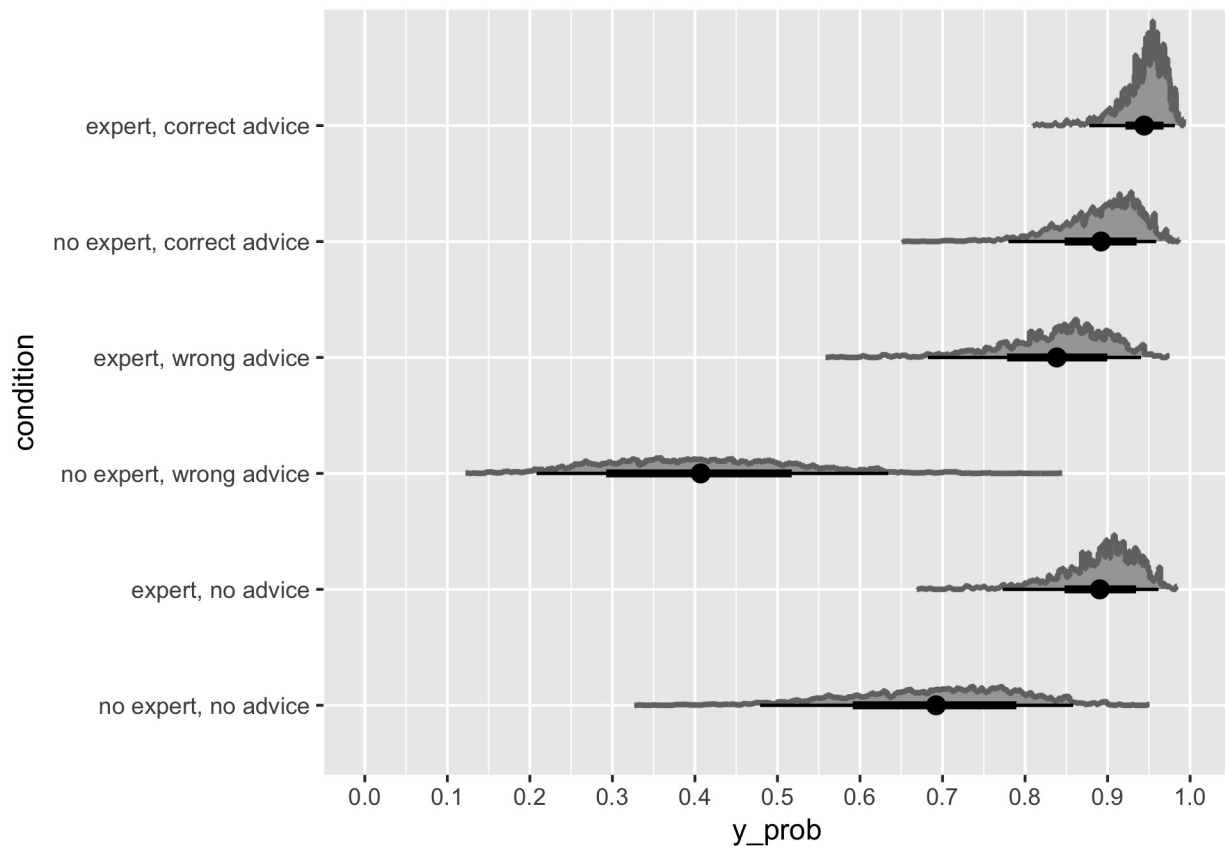


Figure 2. Marginal distributions including means, 66% and 95% confidence intervals for all experimental conditions while setting the standard deviation of item random intercepts to 0.01.

As an example, Figure 2 reveals a number of implicit assumptions about the comparison between experts and non-experts: With wrong advice, virtually all experts have



a higher probability of making a correct diagnosis compared to non-experts when considering only items with average difficulty. In contrast, there is considerable overlap in probability between experts and non-experts with no advice and even higher overlap with correct advice. Patterns like these should be considered carefully and discussed with the domain experts. Parameter values ( $\beta$  parameters, and  $\sigma_S$ ) should be adjusted if the implications do not seem reasonable.

We could also have a closer look at variability between items by setting the subject standard deviation to almost zero ( $\sigma_S = 0.01$ ).

The final plot demonstrates that these plots are also useful for spotting standard deviations that are specified too high. For Figure 3, we have set  $\sigma_S = 3$  and  $\sigma_I = 3$ . This implies that in each experimental condition, the probabilities that a subject solves an item are usually close to either 0 or 1, which is not a plausible assumption. However, these high standard deviations do not account for the inherent variability and complexity of human performance. For example, we would expect that a participant with low ability compared to other task experts to solve a difficult item with a probability substantially larger than zero even when presented with wrong advice.

### **Iterative process with domain experts.**

**THEORY.** In our experience, the gathering of domain knowledge by domain experts and the consecutive specification of population parameter values used in data simulation is not a one-time event but rather an iterative process. In a first step, domain experts can be interviewed to “elicit” their domain knowledge about how the future data of the planned study is expected to look like. As most domain experts are no experts in statistical modeling and GLMMs, they often struggle without further guidance to communicate their knowledge in a way that is useful when specifying the parameters for data simulation. For this reason, we suggest that after an initial unstructured interview of domain experts, the analyst which is familiar with the structure of the GLMM under study selects an initial set of insightful

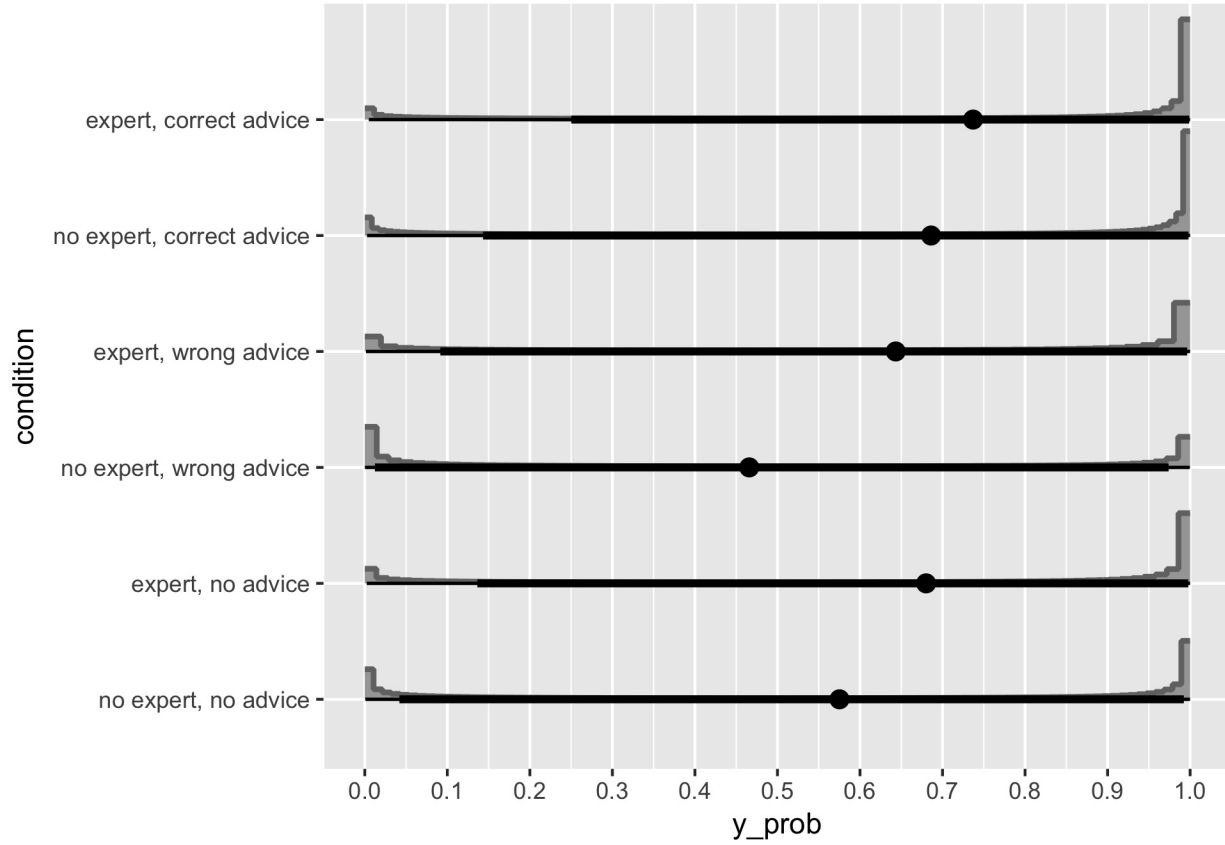


Figure 3. Marginal distributions including means, 66% and 95% confidence intervals for all experimental conditions while setting the standard deviation of subject and item random intercepts to 3.

descriptive statistics and model based quantities. Then they reenter into an iterative discussion where some set of population values are selected and the plausibility of resulting implied quantities are discussed with the domain experts. Then, the population parameters are updated based on this discussion until the domain experts are satisfied with the final result. During this process, the monitored model based quantities and descriptive statistics can be updated or extended to capture as much available domain knowledge as possible.

*PRACTICE.* All parameter values in our present case study have been determined based on repeated discussions with our affiliated domain experts in Radiology to validate our assumptions. Initially, we reviewed the literature to establish a reasonable baseline

performance rate for examining head CT scans for intracranial hemorrhage. Existing studies indicate that radiologists typically demonstrate high accuracies, often exceeding or hovering around 90%, while interns have been shown to perform below 80%, and medical students fall even shorter. For simplicity, we assumed plausible probability values of .90 for experts and .70 for non-experts, respectively. Our experts confirmed that these values are realistic baselines for reviewing diverse head CT images without AI assistance. Subsequently, we consulted several published papers investigating the effect of correct and incorrect advice on decision-making performance in other settings. From their findings, we inferred that both experts and non-experts should benefit from correct and suffer losses from incorrect advice. However, the magnitude of these effects should be substantially greater for non-experts, given their demonstrated reliance on advice compared to experts. We further validated the plausibility of our estimated gains and losses with the collaborating radiologists. For our simulation, we used the probabilities of average participants to solve an average case, as shown in Table 1.

### **Estimate the statistical model**

**THEORY.** At this point, the researcher is capable of producing a simulated dataset similar to the actual dataset that will later be collected in the planned study. The next step is to specify how the statistical model shall be estimated in the actual study collected later. This usually includes the selection of 1) a statistical framework (e.g., frequentist statistics), 2) a software package that is capable of estimating the model class of interest (e.g., the lme4 R package), 3) an estimation algorithm (e.g., the default optimizer “bobyqa”), and 4) the specific model structure including all fixed effects, random effects, and the model family of the dependent variable.

Note that this does not always mean that one will specify the same GLMM that was used when specifying the data generating process. On the one hand, using a simpler model for data simulation than for model estimation can be a useful strategy in scenarios where

making plausible assumptions for complicated random effect structures and interactions is not feasible. On the other hand, using a more complex model (that perhaps even goes beyond a standard GLMM) for data simulation than for model estimation can be a useful strategy in scenarios where one has specific domain knowledge about aspects of the data generating process that are still difficult to estimate with the current state-of-the-art in multilevel modeling.

**PRACTICE.** In our case study, we use the `lme4` R package (Bates, Mächler, Bolker, & Walker, 2015), which is a state-of-the-art tool for fitting frequentist GLMMs.<sup>1</sup> The `lme4` package includes a function called `simulate` that allows researchers to simulate the dependent variable based on the same model formula used for model fitting, enabling simulation-based power analyses and other related analyses.

However, the model parameterization used by the `lme4` package is quite technical, making it difficult for applied researchers to determine whether their specified population model (i.e., the theoretical model that describes the underlying data generation process for a specific population of interest) implies plausible associations in their simulated data. Therefore, for our case study, we have written our own function to simulate data for GLMMs from first principles (i.e., creating synthetic data step by step instead of using black box functions). We think this choice assists applied researchers in better understanding all model assumptions before using `lme4` to analyze the simulated datasets.<sup>2</sup>

In this section, we show how to fit a GLMM with `lme4`. We simulate data according to our model, in which 100 subjects respond to 50 items (we use `set.seed` to make the simulation reproducible). However, for the sake of the exercise, we can imagine that this would be real data resulting from our future experiment and think about how we would

---

<sup>1</sup> For Bayesian GLMMs, the `brms` R package is currently the most prominent option (Bürkner, 2017).

<sup>2</sup> A less flexible alternative would be to use the `simr` package (Green & MacLeod, 2016), which can be used to both simulate data and perform power analysis for models supported by the `lme4` package.

analyze this data.

```
library(tidyverse)

set.seed(1)

dat <- simulate(n_subjects = 100, n_items = 50)
```

The lme4 package uses a special syntax for model specification. Our specific GLMM is represented by the formula:

```
library(lme4)

f <- y_bin ~ 1 + expert + advice_present + advice_correct +
  expert:advice_present + expert:advice_correct +
  (1|subject) + (1|item)
```

The first two lines look similar to any linear model in R (general intercept indicated by 1; main effects indicated by variable names in the dataset; interactions indicated by variable1:variable2). The third line specifies a random intercept for each subject (1|subject) and for each item (1|item). The complete set of rules for the syntax is outlined in Bates et al. (2015) and in the documentation of the lme4 package.

In lme4, a GLMM is fitted with the `glmer` function. By setting `family = "binomial"`, we request a binomial GLMM appropriate for our binary dependent variable `y_bin` (the binomial GLMM uses the canonical logit link by default), which is defined as an accurate (1) vs. inaccurate (0) diagnosis.

```
fit <- glmer(f, data = dat, family = "binomial")
```

We can inspect the estimates for all model parameters with the `summary` command:

```
summary(fit)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
```

```
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula:
## y_bin ~ 1 + expert + advice_present + advice_correct + expert:advice_present +
## expert:advice_correct + (1 | subject) + (1 | item)
## Data: dat
##
##      AIC      BIC   logLik deviance df.resid
##  4149.4   4201.6  -2066.7   4133.4     4992
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -5.7669  0.2125  0.3046  0.4317  2.1056
##
## Random effects:
##  Groups  Name          Variance Std.Dev.
##  subject (Intercept) 0.3148   0.5611
##  item     (Intercept) 0.1624   0.4029
## Number of obs: 5000, groups:  subject, 100; item, 50
##
## Fixed effects:
##
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.0339    0.1103   9.374 < 2e-16 ***
## expert           1.1849    0.2096   5.654 1.57e-08 ***
## advice_present   -1.3436    0.1206 -11.143 < 2e-16 ***
## advice_correct    2.6154    0.1273  20.540 < 2e-16 ***
## expert:advice_present 1.0589    0.2940   3.601 0.000317 ***
```

```
## expert:advice_correct  -1.8104      0.2915  -6.210 5.28e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##              (Intr) expert advc_p advc_c exprt:dvc_p
## expert      -0.377
## advic_prsnt -0.349  0.176
## advic_crrct  0.023  0.001 -0.668
## exprt:dvc_p  0.143 -0.448 -0.412  0.276
## exprt:dvc_c -0.008  0.004  0.292 -0.435 -0.686
```

In the lme4 output, the `Estimate` column in the `Fixed effects` table contains the estimates for the  $\beta$  parameters, while the `Std.Dev.` column in the `Random effects` table contains the estimates for  $\sigma_S$  and  $\sigma_I$ .

### Compute the estimate

**THEORY.** When the theoretical and empirical estimands have been identified in the modeling framework used to estimate the GLMM, one has to decide on whether hypothesis testing or interval estimation should be used to answer the research question. While the majority of psychological research uses hypothesis testing, there are repeated claims that a focus on estimation can also be a good choice in many circumstances (CITATION). This decision on testing or estimating is then followed by selecting the specific statistical method that shall be applied to compute the hypothesis test(s) or confidence interval(s) (e.g., compute hypothesis tests and confidence intervals with the `marginaleffects` R package using the delta method).

**PRACTICE.** For our case study, we investigate the following research question: *We expect that for both experts and non-experts, correct advice leads to a higher probability of*

*accurately diagnosing a CT scan compared to no advice presented, AND, we expect that for both experts and non-experts, incorrect advice leads to a lower probability of accurately diagnosing a CT scan compared to no advice presented.*

In the earlier section on our theoretical estimand we translated this verbal description into four probability statements that are specified outside of any specific statistical model. The next translation step is to specify the empirical estimand, that is how these probabilities can be estimated within the scope of our concrete GLMM.

The first of the four statements of the theoretical estimand was:

$$\begin{aligned} &P(\text{correct diagnosis}|\text{correct advice, average expert, average scan}) \\ &- P(\text{correct diagnosis}|\text{no advice, average expert, average scan}) \end{aligned}$$

In the context of our GLMM, the corresponding probability statement is:

$$\begin{aligned} &P(Y = 1|advice\_present = 1, advice\_correct = 1, expert = 1, u_{0s} = 0, u_{0i} = 0) \\ &- P(Y = 1|advice\_present = 0, advice\_correct = 0, expert = 1, u_{0s} = 0, u_{0i} = 0) \end{aligned}$$

Plugging in the model equation of the GLMM produces an equation on how to compute this contrast when all model parameters are known. We have already discussed how to compute the individual probabilities in the section on specifying population parameters. When we want to estimate the above contrast based on observed data, the only difference is that model parameters are not known and we instead plug the corresponding parameter estimates into the equation.

We could use our knowledge of the structure of our GLMM to determine the exact formula needed to compute our contrast of interest and then plug in the parameter estimates manually from the `summary(fit)` output. However, this would be very tedious and we can use R to compute this contrast without doing the math.



We could use the `predict` function of the `lme4` package to compute the predicted probability for a correct diagnosis based on our model, plug in the two sets of predictor values, and compute the difference between the two estimated probabilities.

```
pred_df <- data.frame(advice_present = c(1, 0), advice_correct = c(1, 0),
  expert = c(1, 1))
pred_df
```

```
##  advice_present advice_correct expert
## 1             1             1      1
## 2             0             0      1
```

```
preds <- predict(fit, newdata = pred_df, type = "response", re.form = NA)
preds
```

```
##           1           2
## 0.939292 0.901923
```

```
preds[1] - preds[2]
```

```
##           1
## 0.03736901
```

The argument `type = "response"` specifies that predictions are made on the probability scale (instead of the log-odds scale of the  $\beta$  parameters), while `re.form = NA` sets all random effects to 0.

We could use this method to compute point estimates for all four contrast that are part of our empirical estimand. However, depending on whether we are interested in hypothesis testing or parameter estimation, we also need a method to compute hypothesis tests or confidence interval for our point estimates.

The `marginalEffects` package (Arel-Bundock, 2023) is a very flexible, increasingly popular package to compute hypothesis tests and confidence intervals for contrasts with a variety of statistical models, including GLMMs estimated with `lme4`.

First, we specify a grid of all combinations of predictor variable and then compute estimated probabilities for all experimental conditions in our experiment with the `predictions` function:

```
pred_df <- expand_grid(advice_present = 0:1, advice_correct = 0:1, expert = 0:1)
pred_df
```

```
## # A tibble: 8 x 3
##   advice_present advice_correct expert
##           <int>         <int> <int>
## 1             0             0     0
## 2             0             0     1
## 3             0             1     0
## 4             0             1     1
## 5             1             0     0
## 6             1             0     1
## 7             1             1     0
## 8             1             1     1
```

```
probs <- predictions(fit, newdata = pred_df, type = "response", re.form = NA)
probs
```

```
##
##   Estimate Std. Error      z Pr(>|z|)      S 2.5 % 97.5 % advice_present
##     0.738    0.02134   34.6  <0.001 867.2 0.696  0.779             0
##     0.902    0.01739   51.9  <0.001  Inf 0.868  0.936             0
```

```
##      0.975      0.00421 231.6    <0.001    Inf 0.966  0.983          0
##      0.954      0.01454  65.6    <0.001    Inf 0.925  0.982          0
##      0.423      0.03221  13.1    <0.001 128.6 0.360  0.486          1
##      0.874      0.02793  31.3    <0.001 711.4 0.819  0.928          1
##      0.909      0.00967  94.1    <0.001    Inf 0.890  0.928          1
##      0.939      0.01091  86.1    <0.001    Inf 0.918  0.961          1
##  advice_correct expert
##              0      0
##              0      1
##              1      0
##              1      1
##              0      0
##              0      1
##              1      0
##              1      1
##
## Columns: rowid, estimate, std.error, statistic, p.value, s.value, conf.low, conf.high
## Type:  response
```

The point estimates for all experimental conditions are reported in the **Estimate** column. Note that the output also contains the two missing by design conditions that will never be observed in the actual study ( $advice\_present = 0, advice\_correct = 1, expert = 1$  and  $advice\_present = 0, advice\_correct = 1, expert = 0$ ). This is no problem as long as we never interpret those estimates.

Next, we use the estimated probabilities to compute the four specific contrasts that are part of our empirical estimand. For this we must specify which rows in **probs** have to be subtracted from each other. We will use the **hypotheses** function to compute our four

contrasts of interest together with hypothesis tests and confidence intervals. We use the default inference options of the `marginalEffects` package that compute hypothesis tests and confidence intervals based on the approximate delta method.

```
probs %>%
```

```
  hypotheses(hypothesis = c(
    "b8 = b2",
    "b2 = b6",
    "b7 = b1",
    "b1 = b5"),
  equivalence = c(0, 0))
```

```
##
```

```
##   Term Estimate Std. Error      z Pr(>|z|)      S    2.5 % 97.5 % p (NonSup)
## b8=b2    0.0374     0.0162   2.31   0.021   5.6  0.00563 0.0691     0.989
## b2=b6    0.0282     0.0279   1.01   0.312   1.7 -0.02653 0.0830     0.844
## b7=b1    0.1717     0.0173   9.93  <0.001  74.8  0.13780 0.2056     1.000
## b1=b5    0.3145     0.0280  11.24  <0.001  95.0  0.25965 0.3693     1.000
```

```
##   p (NonInf) p (Equiv)
```

```
##         0.0105      0.989
```

```
##         0.1562      0.844
```

```
##         <0.001      1.000
```

```
##         <0.001      1.000
```

```
##
```

```
## Columns: term, estimate, std.error, statistic, p.value, s.value, conf.low, conf.high,
```

```
## Type: response
```

The expression `"b8 = b2"` is special syntax to subtract the estimate in row number 8 from the estimate in row number 2. The argument `equivalence = c(0, 0)` can be used to

compute one-sided p-values, testing whether the contrast in the population is smaller than 0 (**p (NonSub)** column) or greater than 0 (**p (NonInf)** column). The point estimates for four contrasts are reported in the **Estimate** column. Note that to facilitate interpretation, we arranged the contrasts in a way that we theoretically expect positive values for all four of them.

**Hypothesis testing.** If we chose hypothesis testing for our case study, we would test a combined null hypothesis  $H_0$  that consists of four separate null hypotheses:

$$H_{01} : P(Y = 1 | \text{advice\_present} = 1, \text{advice\_correct} = 1, \text{expert} = 1, u_{0s} = 0, u_{0i} = 0) \leq$$

$$P(Y = 1 | \text{advice\_present} = 0, \text{advice\_correct} = 0, \text{expert} = 1, u_{0s} = 0, u_{0i} = 0)$$

$$H_{02} : P(Y = 1 | \text{advice\_present} = 0, \text{advice\_correct} = 0, \text{expert} = 1, u_{0s} = 0, u_{0i} = 0) \leq$$

$$P(Y = 1 | \text{advice\_present} = 1, \text{advice\_correct} = 0, \text{expert} = 1, u_{0s} = 0, u_{0i} = 0)$$

$$H_{03} : P(Y = 1 | \text{advice\_present} = 1, \text{advice\_correct} = 1, \text{expert} = 0, u_{0s} = 0, u_{0i} = 0) \leq$$

$$P(Y = 1 | \text{advice\_present} = 0, \text{advice\_correct} = 0, \text{expert} = 0, u_{0s} = 0, u_{0i} = 0)$$

$$H_{04} : P(Y = 1 | \text{advice\_present} = 0, \text{advice\_correct} = 0, \text{expert} = 0, u_{0s} = 0, u_{0i} = 0) \leq$$

$$P(Y = 1 | \text{advice\_present} = 1, \text{advice\_correct} = 0, \text{expert} = 0, u_{0s} = 0, u_{0i} = 0)$$

The combined null hypothesis  $H_0$  should only be rejected if **all** individual null hypotheses are rejected (i.e., intersection-union setting; Dmitrienko & D’Agostino, 2013). In such cases, the error probabilities do not accumulate, and we would waste power when correcting for multiple tests.

With a standard significance level of  $\alpha = 0.05$ , we would not reject all four null hypotheses (the p-value in the **p (NonInf)** column for the third hypothesis is not significant) and therefore also not reject the combined null hypothesis for this particular (simulated) dataset. Note that this decision would be wrong because we have simulated the data such that the combined alternative hypothesis  $H_1$  is actually true in the population.

**Interval estimation.** If we chose parameter estimation for our case study, we would focus on the two-sided confidence intervals of the four contrasts of interest. With a standard confidence level of  $1 - \alpha = 0.95$ , plausible values are clearly in the positive range for the first, third and fourth contrast, while both negative and positive values seem plausible for the second contrast. Note that due to the constrained range of the probability scale, the width of the confidence interval differs between the four contrasts (which is the expected behavior for binomial GLMMs). The smallest width is observed for the first contrast (expert with correct advice vs. expert without advice) where both underlying probabilities are close to 1. The largest width is observed for the fourth contrast (non-expert with wrong advice vs. non-expert without advice), where both underlying probabilities are closer to 0.5.

### Perform repeated simulations

**THEORY.** Conducting all previous steps enables the analyst to 1) simulate a dataset, 2) estimate a GLMM, and 3) compute hypothesis tests or confidence intervals for estimands of interest, mirroring the analysis that will later be performed for the actual dataset of the planned study. The last missing piece is to write code to perform the above steps repeatedly and allow for a setting using different sample sizes. On a conceptual level, we first require a function, that takes as input the sample size and the full set of population parameter values. When planning for power, the function should return the p-value(s) of the hypothesis test(s) of interest when conducted on the simulated dataset. When planning for precision, the function should return the width of the confidence interval(s) of interest. Secondly, we must run this function repeatedly with the same sample size and population parameters. Because even fitting GLMMs with frequentist methods can quickly become time-consuming, it is recommended to use parallel computing, that is running simulations on multiple cores of the computer at the same time to reduce total run time. Thirdly, the results of the repeated simulation must be collected and aggregated. When planning for power, we compute the relative frequency of a significant p-value(s) across repeated

simulations. When planning for precision, we compute the average width of the confidence interval(s). Lastly, we have to repeat the complete simulation for different sample sizes, to determine how big the sample must be in order to achieve the targeted power or precision.

**PRACTICE.** We are finally ready to run our simulation-based sample size planning analyses to plan for power and for precision. Wrapping the `simulate` function already constructed earlier, the helper function `sim_and_analyse` performs all previous steps (simulate a dataset, fit a GLMM, compute p-values and confidence intervals) in a single command.

```
sim_and_analyse <- function(
  formula_chr = "y_bin ~ 1 + expert + advice_present + advice_correct +
    expert:advice_present + expert:advice_correct + (1|subject) + (1|item)",
  contrasts = c("b8 = b2", "b2 = b6", "b7 = b1", "b1 = b5"), ...){
  require(lme4)
  require(marginaleffects)
  require(tidyr)
  # simulate data
  dat <- simulate(...)
  # fit model
  model <- glmer(as.formula(formula_chr), data = dat, family = "binomial")
  # compute contrasts
  contr_df <- expand_grid(advice_present = 0:1, advice_correct = 0:1,
    expert = 0:1)
  predictions(model, newdata = contr_df, type = "response", re.form = NA) %>%
    hypotheses(hypothesis = contrasts, equivalence = c(0, 0)) %>%
    data.frame()
}
```

Simulation-based sample size planning can quickly become computationally intensive when we repeatedly simulate data and fit models for different parameter combinations or sample sizes. Thus, we use the `future` (Bengtsson, 2021) and `furrr` (Vaughan & Dancho, 2022) packages to perform computations in parallel. First, we enable parallelization with the `plan` function and specify how many parallel cores (“workers”) of our computer to use (users can find out the maximum number of cores on their computer with the command `parallel::detectCores()`), and set a seed to make the simulation reproducible.

```
library(future)
plan("multisession", workers = 6)
set.seed(2)
```

The next code chunk specifies a simulation grid with different settings for both the number of subjects (`n_subjects`) and the number of items (`n_items`), each combination being repeated `rep` times. We chose 300 repetitions for the data simulation at hand as it strikes a balance between achieving a robust statistical estimate and remaining computationally feasible. With the current settings, this simulation takes about one hour on a MacBook Pro from 2020 with M1 chip and 16 GB working memory. If you want to quickly experiment with the code yourself, a setting with `workers = 4` and `rep = 5` should finish in less than 5 minutes, even on smaller machines.

```
library(furrr)
sim_result <- crossing(
  rep = 1:300,
  n_subjects = c(100, 150, 200, 250),
  n_items = c(10, 30, 50, 70)
) %>%
  mutate(res = future_pmap(., sim_and_analyse,
    .options = furrr_options(seed = TRUE))) %>%
```



```
unnest(col = res)
```

The result of this computation is a data frame that contains the p-values and confidence intervals of all specified contrasts for each simulated dataset. In some iterations (predominantly in conditions with small sample sizes), model estimation did not converge with the lme4 package. When the model fails to converge, it means that the statistical model being fitted to the data failed to reach a stable or valid solution during the estimation process. We do not remove these results because non-convergence can also happen when analyzing the real data we plan to collect, thus, we want to factor in this possibility to keep our simulation more realistic.

**Power results.** For our exemplary combined hypothesis, power is defined as the (long-run) percentage of simulations in which all four p-values of our individual hypotheses are significant at the  $\alpha = 0.05$  level. Based on our simulation outcomes, we compute a power estimate for each combination of `n_subjects`  $\times$  `n_items` (including 95% confidence intervals) and visualize the results with the following code.<sup>3</sup>

```
library(binom)
alpha <- 0.05
power <- sim_result %>%
  pivot_wider(names_from = term, names_sep = "_",
    values_from = estimate:p.value.equiv) %>%
  group_by(n_subjects, n_items) %>%
  summarise(
    power = mean(`p.value.noninf_b1=b5` < alpha &
      `p.value.noninf_b8=b2` < alpha & `p.value.noninf_b2=b6` < alpha &
      `p.value.noninf_b7=b1` < alpha),
```

---

<sup>3</sup> This code was inspired by the “Mixed Design Simulation” vignette of the faux package at [https://debruine.github.io/faux/articles/sim\\_mixed.html](https://debruine.github.io/faux/articles/sim_mixed.html).

```

n_sig = sum(`p.value.noninf_b1=b5` < alpha &
            `p.value.noninf_b8=b2` < alpha & `p.value.noninf_b2=b6` < alpha &
            `p.value.noninf_b7=b1` < alpha),

n = n(),

ci.lwr = binom.confint(n_sig, n, method = "wilson")$lower,
ci.upr = binom.confint(n_sig, n, method = "wilson")$upper,

.groups = "drop")

power %>%

mutate(across(c(n_subjects, n_items), factor)) %>%

ggplot(aes(n_subjects, n_items, fill = power)) +

geom_tile() +

geom_text(aes(label = sprintf("%.2f \n [%.2f; %.2f]",
                                power, ci.lwr, ci.upr)),

          color = "white", size = 4) +

scale_fill_viridis_c(limits = c(0, 1)) +

xlab("number of subjects") + ylab("number of items")

```

As should be the case, power estimates in Figure 4 increase with both the number of subjects and the number of items. The confidence intervals reported here indicate how precisely power was estimated by our simulation. Higher precision (which would be reflected in narrower confidence intervals) could be obtained by increasing the number of repetitions (`rep`) in the simulation. In practice, data simulations are often run multiple times with adjusted combinations of sample sizes. When running for the first time, it might be revealed that power is way too low (or much higher than required) for some combinations of `n_subjects` and `n_items`. When narrowing down the best combination that achieves sufficient power while at the same time striking a good balance of how many subjects and items are practically feasible, later rounds of data simulation will typically include a smaller

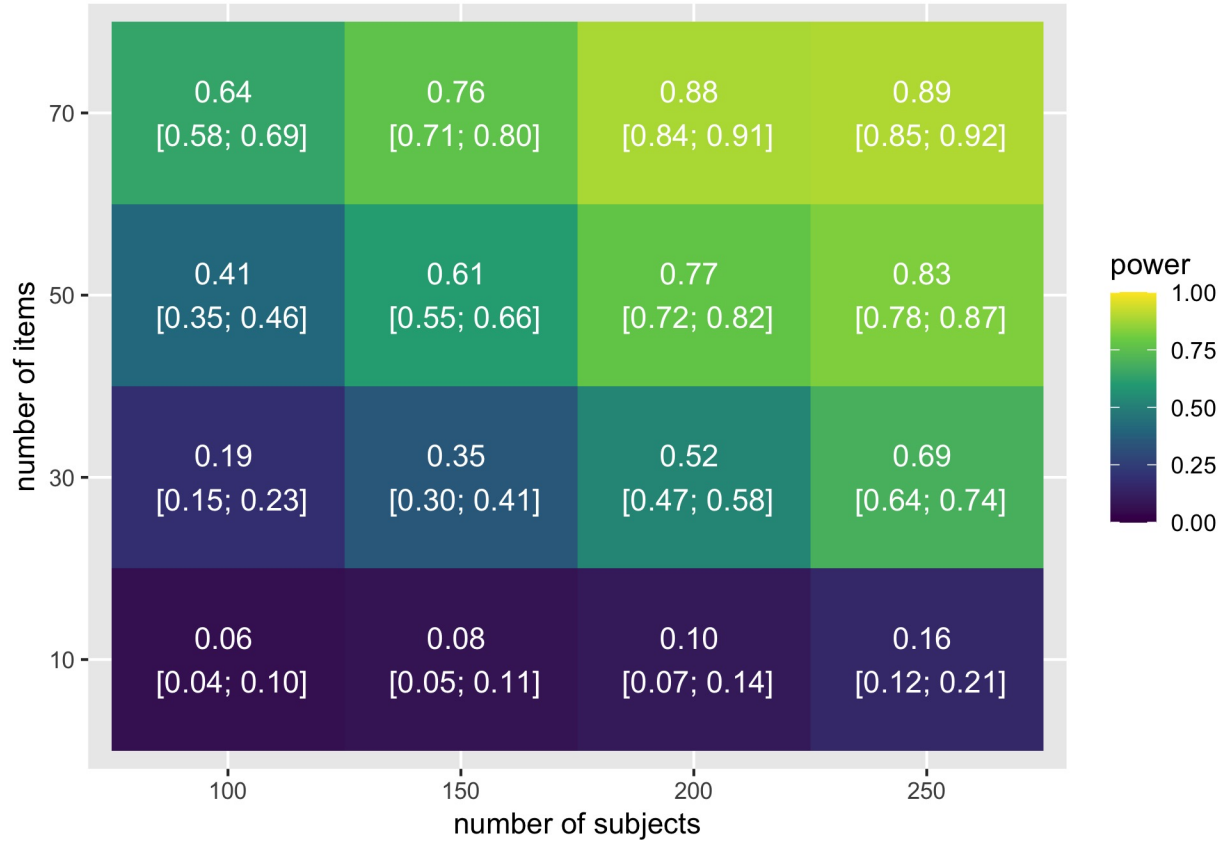


Figure 4. Simulation-based power estimates including 95% confidence interval of the case study for different numbers of subjects and items, based on a significance level of 0.05.

grid of sample sizes combined with a higher number of repetitions. This will assure high precision for the final power estimates, which are then used for the sample size justification of the future study.

Much has been written on the optimal amount of power to target in empirical research. The most prominent heuristic is to target a power of 0.8 (when combined with a type I error rate of  $\alpha = 0.05$ ), but depending on the research goals of the study, there are often good reasons to move away from this standard depending on the research goals and resource constraints (Lakens, 2022b; Lakens et al., 2018). When target power has been specified, the number of subjects and the number of items in our study design can be traded against each other based on practical considerations. For the sake of the example, let the targeted power

be indeed about 0.8, using an  $\alpha$  of 0.05 to detect an effect of the expected size implied by our data simulation. This could be achieved by collecting data from 200 subjects (about 25% of which will be experts), each completing the same 50 items (with advice present in about 67% of cases, which is correct in about 80% of cases with present advice). If collecting data from 200 subjects is not feasible, an alternative would be to recruit 150 subjects but increase the length of the experiment to over 70 items. However, 70 items might take too long to complete for the radiologists participating in the study, who have a busy schedule. The simulation suggests that it might also be possible to plan a shorter experiment with only 30 items if it is feasible to recruit an even higher number of subjects ( $> 250$ , to be determined by additional rounds of power analysis). Design parameters that also affect power, and which could be investigated in the simulation to find a more optimal trade-off, are the ratio of experts, the frequency of whether advice is presented and whether it is correct.

***Precision results.*** When planning for precision, one could monitor the width of all four confidence intervals at the same time. However, because the confidence intervals of the four contrasts strongly differ in width, it is not trivial to decide which width one should target when deciding on the appropriate sample size. In contrast to planning for power, there are no common standards on how to specify the targeted precision. For our example, we use a simple heuristic but we strongly encourage readers to think about better alternatives that are appropriate in their own applications.

Our simulations show that the smallest confidence interval can be expected for the first contrast (expert with correct advice vs. expert without advice). The true contrast in probability for an average expert and an average item in this condition is  $\text{plogis}(b_0 + b_e + b_a + b_c + b_{ea} + b_{ec}) - \text{plogis}(b_0 + b_e) = 0.05$ . We want the width of this confidence interval to be smaller than 0.1. This would mean that if the point estimate happens to be close to the true value, the plausible values inside of a 95% confidence interval would all be positive.

Thus in our example, precision is defined as the (long-run) average width of a 95% confidence interval for the probability contrast between experts with correct advice and experts without advice. Of course, lower width implies better precision. Based on our simulation outcomes, we compute the precision estimate for each combination of  $n\_subjects \times n\_items$  (including 95% confidence intervals) and visualize the results with the following code.

```
precision <- sim_result %>%
  pivot_wider(names_from = term, names_sep = "_",
    values_from = estimate:p.value-equiv) %>%
  group_by(n_subjects, n_items) %>%
  mutate(width = `conf.high_b8=b2` - `conf.low_b8=b2`) %>%
  summarise(precision = mean(width),
    ci.lwr = t.test(width)$conf.int[1],
    ci.upr = t.test(width)$conf.int[2],
    .groups = "drop")
precision %>%
  mutate(across(c(n_subjects, n_items), factor)) %>%
  ggplot(aes(n_subjects, n_items, fill = precision)) +
  geom_tile() +
  geom_text(aes(label = sprintf("%.2f \n [%.2f; %.2f]",
    precision, ci.lwr, ci.upr)),
    color = "white", size = 4) +
  scale_fill_viridis_c(limits = c(0, 0.3), direction = -1) +
  guides(fill = guide_legend(reverse=FALSE))
```

As should be the case, precision estimates in Figure 5 increase (i.e., average width of confidence interval decreases) with both the number of subjects and the number of items.

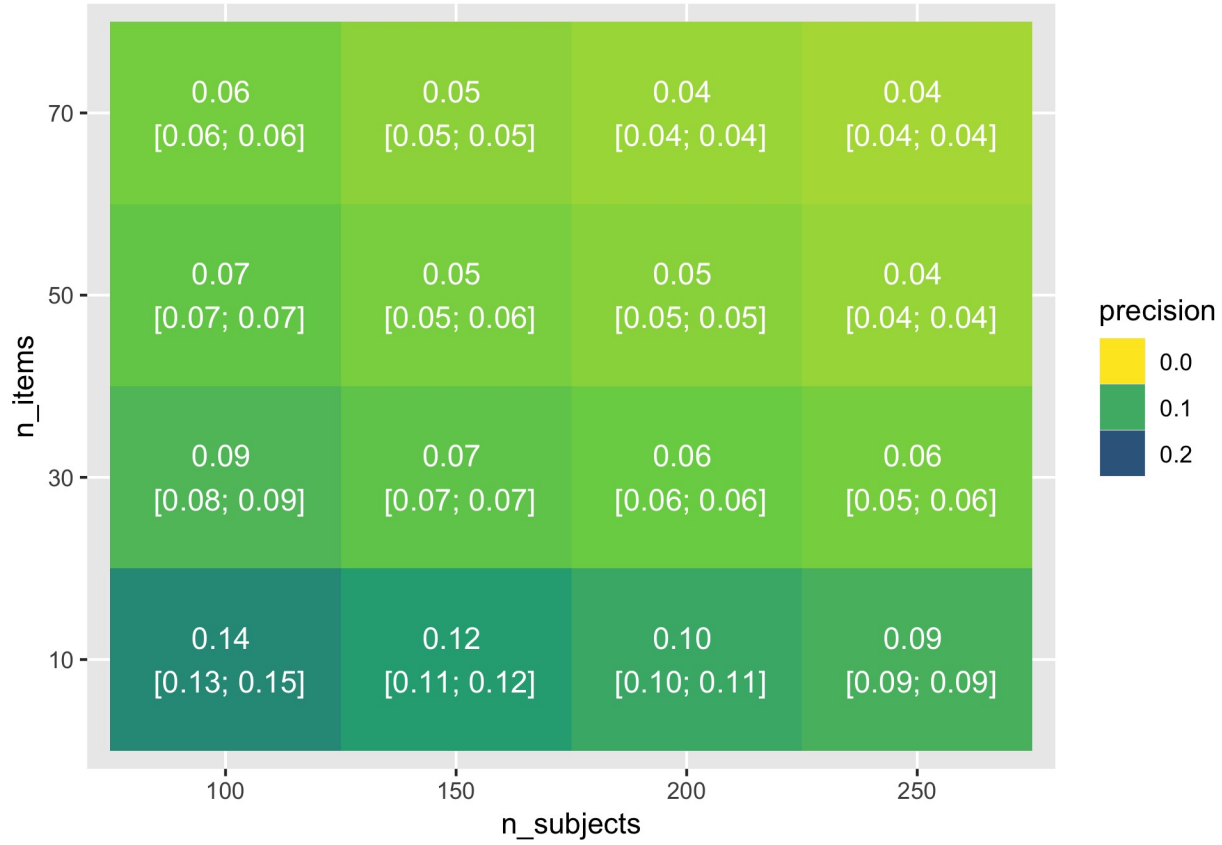


Figure 5. Simulation-based precision estimates (expected width of confidence intervals) including 95% confidence interval of the case study for different numbers of subjects and items, based on a confidence level of 0.95.

The confidence intervals reported here indicate how precisely the expected width of the confidence interval for our focal contrast was estimated by our simulation. Applying our simple heuristic of targeting an expected width smaller than 0.1, we see the same trade-off between the number of subjects and the number of items as with planning for power. We could either choose 100 subjects and 30 items or 200 subjects and 10 items.

Note that our simple heuristic for determining sample size in the planning for precision scenario was quite liberal. This is reflected by the result that we would need a smaller sample size than in the planning for power scenario. With a more conservative precision target, the result is generally the opposite: As a rule, precise parameter estimates usually

require bigger samples than null hypothesis testing.

## Sensitivity analysis

In our case study, we have performed simulation-based sample size planning from a single set of parameter values that reflect our assumptions of an expected effect size. Instead of extracting this expected effect size from meta-analyses or pilot data, which has been the main focus of previous tutorials, we have demonstrated some strategies to determine plausible parameter values in GLMMs based on domain knowledge. Domain knowledge can be considered a vague theoretical model about the data-generating process that is less formal and can only be accessed by a back-and-forth exchange in which domain experts assess the plausibility of simulated data. When sample sizes are chosen based on the results of our simulation-based power analysis, a future study will be informative to reject the null hypothesis if an effect of our *expected size* is present (or estimate the effect with satisfying precision). However, if the true effect is indeed smaller, power (or precision) will be lower, and the study might not be sufficiently informative. A common, more conservative strategy for sample size justification is to perform sample size planning for the smallest effect size of interest (SESOI). An effect smaller than the SESOI would be considered too small to be interesting or practically meaningful, even if the effect is not actually zero (King, 2011). For strategies on the even more difficult task of specifying a plausible SESOI, as well as a thorough discussion of various topics concerning power analysis, see (Lakens, 2022a). When domain knowledge or formal theories about the research topic of interest are too vague to specify a meaningful SESOI, it is still recommended to demonstrate power or precision for different effect sizes in what is called *sensitivity power analysis*. By simulating power (or precision) for different effect sizes (in addition to the different number of subjects and items), one can make sure that power (or precision) would still be sufficient to detect smaller effect sizes than our expected effect or at least get an impression of how strongly power (or precision) depends on the size of the true effect. In simple study designs, it is possible to

perform sensitivity analysis based on a single standardized effect size (e.g., analyze power in a two-sample t-test for a standardized mean difference varying between 0.1 and 0.8).

However, for our case study that investigates combined hypotheses in a GLMM modeling framework, the effect size is implicitly represented by the complex distribution of probabilities within and between experimental conditions. In this setting, sensitivity analysis would require manually specifying additional sets of plausible parameter values that reflect scenarios with smaller or larger differences between groups with respect to our specific research question. Power (or precision) could then be simulated for several of these scenarios (across different numbers of subjects and items, as considered earlier).

## Outlook

Beyond the specifics of our concrete case study, we want to outline six developments regarding the future role of simulation-based sample size planning in experimental research:

1. The growing need for simulation-based sample size planning in experimental research:

In order to conduct informative research using varying complex experimental designs with GLMMs, formula-based heuristics and user-friendly software tools for a priori power analysis are often not suitable. Therefore, simulation-based power analysis is becoming increasingly needed since it provides experimental researchers with a tailored approach to estimating required sample sizes before data collection.

2. Managing data simulations more easily with discrete predictor variables:

Simulation-based sample size planning becomes more manageable when all predictor variables are discrete (like in the presented case study) and fixed by the study design. This allows researchers to focus on simulating outcome variables while avoiding the need for complex simulations of predictor values, which would introduce additional assumptions. By simplifying the simulation process, researchers can obtain reliable estimates for power or precision without compromising realistic assumptions about the



data-generating process implied by the study design.

3. Teaching data simulation skills: The ability to conduct simulation-based sample size planning is a valuable skill that should be taught to experimental researchers. By incorporating such training into research methods courses and workshops, researchers can gain a deeper understanding of statistical power or precision, and improve the quality of their experimental designs. Equipping researchers with the knowledge and tools to perform simulation-based sample size planning enables them to make informed decisions and enhance the rigor of their studies. The need to reason about how to simulate plausible data that is in line with the research hypothesis, while not violating domain expertise on how plausible data should look, might also contribute to planning more insightful studies that can answer more precise research questions (Yarkoni, 2022).
4. Addressing the mismatch in effort perception: There is often a significant disconnect between the amount of effort required to perform a priori simulation-based sample size planning and the perceived effort estimated by researchers and collaborators in experimental research. Many researchers request simulation-based power analyses from statisticians or methodological experts without fully comprehending the complexity and time-consuming nature of these tailored simulations. It is crucial to raise awareness about the effort involved to ensure realistic expectations and effective collaboration between researchers and methodological experts.
5. Recognizing the value of simulation-based design analysis: Simulation-based power analyses are not mere technicalities; they are valuable research contributions that deserve recognition in experimental research. They offer insights into the robustness and sensitivity of experimental designs, helping researchers make informed decisions about sample sizes, effect sizes, and statistical power or precision. Their importance can be reflected by allocating them a separate publication or incorporating them as a

significant component of stage 1 preregistered reports (Chambers & Tzavella, 2022).

6. Integration with Open Science and preregistration practices: Simulation-based sample size planning aligns well with the principles of Open Science and preregistration in experimental research. When researchers have access to simulated data based on their pre-specified model, analyzing the collected dataset becomes straightforward and unambiguous. By preregistering their simulation-based sample size plan, researchers enhance the transparency and accountability of their experimental procedures, contributing to the credibility and reproducibility of research.

## Conclusion

In the wake of the replication crisis and myriad of underpowered experimental work, generalized linear mixed models (GLMMs) offer a flexible statistical framework to analyze experimental data with complex (e.g., dependent and hierarchical) data structures. Yet, analytic methods and software cannot be applied to conduct a priori sample size planning for GLMMs, necessitating data simulation-based approaches. Through this applied tutorial, we aim to provide researchers with the necessary skills and tools to perform simulation-based sample size planning with GLMMs themselves. By incorporating GLMMs and a priori sample size planning into their work, researchers can enhance the replicability and credibility of their experiments (Yarkoni, 2022).

## References

- Arel-Bundock, V. (2023). *MarginalEffects: Predictions, comparisons, slopes, marginal means, and hypothesis tests*. Retrieved from <https://CRAN.R-project.org/package=marginalEffects>
- Arend, M. G., & Schäfer, T. (2019). Statistical power in two-level models: A tutorial based on Monte Carlo simulation. *Psychological Methods*, 24(1), 1–19. <https://doi.org/10.1037/met0000195>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using **Lme4**. *Journal of Statistical Software*, 67(1). <https://doi.org/10.18637/jss.v067.i01>
- Bengtsson, H. (2021). A unifying framework for parallel and distributed processing in r using futures. *The R Journal*, 13(2), 208–227. <https://doi.org/10.32614/RJ-2021-048>
- Brysbaert, M., & Stevens, M. (2018). Power Analysis and Effect Size in Mixed Effects Models: A Tutorial. *Journal of Cognition*, 1(1), 9. <https://doi.org/10.5334/joc.10>
- Bürkner, P.-C. (2017). Brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, 80, 1–28. <https://doi.org/10.18637/jss.v080.i01>
- Chambers, C. D., & Tzavella, L. (2022). The past, present and future of Registered Reports. *Nature Human Behaviour*, 6(1), 29–42. <https://doi.org/10.1038/s41562-021-01193-7>
- Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., Weibelzahl, S., Anandkumar, A., . . . De Rosario, M. H. (2018). Package “pwr.” *R Package Version*, 1(2).
- DeBruine, L. (2023). *Faux: Simulation for factorial designs*. Zenodo. <https://doi.org/10.5281/zenodo.2669586>
- DeBruine, L., & Barr, D. J. (2021). Understanding Mixed-Effects Models Through Data Simulation. *Advances in Methods and Practices in Psychological Science*, 4(1), 2515245920965119. <https://doi.org/10.1177/2515245920965119>
- Dmitrienko, A., & D’Agostino, R. (2013). Traditional multiplicity adjustment methods in clinical trials. *Statistics in Medicine*, 32(29), 5172–5218.

<https://doi.org/10.1002/sim.5990>

Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. D. (2021). *Regression: Models, Methods and Applications*. Berlin, Heidelberg: Springer Berlin Heidelberg.

<https://doi.org/10.1007/978-3-662-63882-8>

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G\*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(4), 1149–1160. <https://doi.org/10.3758/BRM.41.4.1149>

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., . . .

Modrák, M. (2020). *Bayesian Workflow*. arXiv. Retrieved from

<https://arxiv.org/abs/2011.01808>

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, 7(4), 493–498.

<https://doi.org/10.1111/2041-210X.12504>

Kelley, K., Maxwell, S. E., & Rausch, J. R. (2003). Obtaining Power or Obtaining Precision:

Delineating Methods of Sample-Size Planning. *Evaluation & the Health Professions*,

26(3), 258–287. <https://doi.org/10.1177/0163278703255242>

Kelley, K., & Rausch, J. R. (2006). Sample size planning for the standardized mean

difference: Accuracy in parameter estimation via narrow confidence intervals.

*Psychological Methods*, 11(4), 363–385. <https://doi.org/10.1037/1082-989X.11.4.363>

King, M. T. (2011). A point of minimal important difference (MID): A critique of

terminology and methods. *Expert Review of Pharmacoeconomics & Outcomes Research*,

11(2), 171–184. <https://doi.org/10.1586/erp.11.9>

Kumle, L., Vö, M. L.-H., & Draschkow, D. (2021). Estimating power in (generalized) linear

mixed models: An open introduction and tutorial in R. *Behavior Research Methods*,

53(6), 2528–2543. <https://doi.org/10.3758/s13428-021-01546-0>

Lafit, G., Adolf, J. K., Dejonckheere, E., Myin-Germeys, I., Viechtbauer, W., & Ceulemans,

E. (2021). Selection of the Number of Participants in Intensive Longitudinal Studies: A

- User-Friendly Shiny App and Tutorial for Performing Power Analysis in Multilevel Regression Models That Account for Temporal Dependencies. *Advances in Methods and Practices in Psychological Science*, 4(1), 251524592097873.  
<https://doi.org/10.1177/2515245920978738>
- Lakens, D. (2022a). *Improving Your Statistical Inferences*. Zenodo.  
<https://doi.org/10.5281/ZENODO.6409077>
- Lakens, D. (2022b). Sample Size Justification. *Collabra: Psychology*, 8(1), 33267.  
<https://doi.org/10.1525/collabra.33267>
- Lakens, D., Adolfs, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., . . . Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour*, 2(3), 168–171.  
<https://doi.org/10.1038/s41562-018-0311-x>
- Lundberg, I., Johnson, R., & Stewart, B. M. (2021). What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory. *American Sociological Review*, 86(3), 532–565. <https://doi.org/10.1177/00031224211004187>
- Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample Size Planning for Statistical Power and Accuracy in Parameter Estimation. *Annual Review of Psychology*, 59(1), 537–563. <https://doi.org/10.1146/annurev.psych.59.103006.093735>
- Murayama, K., Usami, S., & Sakaki, M. (2022). Summary-statistics-based power analysis: A new and practical method to determine sample size for mixed-effects modeling. *Psychological Methods*. <https://doi.org/10.1037/met0000330>
- Vaughan, D., & Dancho, M. (2022). *Furrr: Apply mapping functions in parallel using futures*. Retrieved from <https://CRAN.R-project.org/package=furrr>
- 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, 2020–2045. <https://doi.org/10.1037/xge0000014>
- Wickham, H., Çetinkaya-Rundel, M., & Grolemund, G. (2023). *R for data science: Import, tidy, transform, visualize, and model data* (2nd edition). Beijing Boston Farnham

Sebastopol Tokyo: O'Reilly.

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, 45, e1.

<https://doi.org/10.1017/S0140525X20001685>

Zimmer, F., Henninger, M., & Debelak, R. (2022). *Sample Size Planning for Complex Study Designs: A Tutorial for the mlpwr Package*. PsyArXiv.

<https://doi.org/10.31234/osf.io/r9w6t>