Application of Confidence Interval Width Contour Plots in Sample Size Planning for Linear Mixed-Effects Models

Authors: Liu Yue, Xu Lei, Liu Hongyun, Han Yuting, You Xiaofeng, Wan Zhilin, Liu Hongyun

Date: 2023-10-07T00:00:00+00:00

Abstract

Linear mixed-effects models exhibit distinct advantages in analyzing psychological experimental data with nested structures. This article proposes the use of confidence interval width contour plots for sample size planning in such models. Through these contour plots, one can determine the numbers of subjects and trials that simultaneously meet the requirements for statistical power, effect size accuracy, and confidence interval width. By examining two representative models that focus on within-subject experimental effects and the moderating effects of subject variables, two simulation studies were conducted employing a Monte Carlo simulation methodology to investigate the influences of effect size, magnitude of random effects, and type of subject variables on confidence interval width contour plots and sample size planning results.

Full Text

Application of Confidence Interval Width Contour Plots in Sample Size Planning for Linear Mixed-Effects Models

LIU Yue[1], XU Lei[1], LIU Hongyun[2],[3], HAN Yuting[4], YOU Xiaofeng[5], WAN Zhilin[1]

[1] Institute of Brain and Psychological Sciences, Sichuan Normal University, Chengdu 610066, China
[2] Beijing Key Laboratory of Applied Experimental Psychology, Beijing Normal University, Beijing 100875, China
[3] Faculty of Psychology, Beijing Normal University, Beijing 100875, China
[4] School of Psychology, Beijing Language and Culture University, Beijing 100083, China

[5] School of Mathematics and Information Science, Nanchang Normal University, Nanchang 360111, China

Abstract

Linear mixed-effects models (LMEMs) offer distinct advantages for analyzing nested data structures common in psychological experiments. This paper proposes the use of confidence interval (CI) width contour plots for sample size planning in LMEMs. These contour plots enable researchers to determine the optimal combination of number of participants and number of trials that simultaneously satisfy criteria for statistical power, effect size accuracy, and CI width. Focusing on two typical model formulations—one examining within-subject experimental effects and the other examining between-subject moderator effects—we conducted two simulation studies using Monte Carlo methods to explore how effect size, magnitude of random effects, and type of subject variable influence CI width contour plots and sample size planning outcomes.

Keywords: linear mixed-effects models, multilevel models, power analysis, effect size, confidence interval width

Introduction

In recent years, psychological researchers have increasingly discussed research misconduct and reproducibility issues. More academic journals, both domestically and internationally, have adopted pre-registration systems, which effectively prevent problematic practices such as p-hacking and promote transparency in research processes and results, thereby enhancing reproducibility (Nosek et al., 2022). Pre-registration requires explicit planning and justification for design elements such as sample size and number of trials. Consequently, determining how to conduct sample size planning for specific statistical models has become a key concern for psychological researchers. This study addresses sample size planning for LMEMs by exploring a paradigm that combines power analysis and effect size accuracy using simulation methods, and by developing intuitive CI width contour plots to help applied researchers identify appropriate numbers of participants and trials, thereby providing methodological support for research design and quality assurance.

1.1 The Sample Size Planning Problem for Linear Mixed-Effects Models

As research questions deepen and data collection methods advance, designs involving random effects for stimuli and nested structures have become increasingly common. For example, psycholinguistic experiments typically use words as stimuli, but different words elicit different reaction times, meaning that observed experimental effects may be partially attributable to stimulus differences (Barr et al., 2013). Traditional methods such as ANOVA confound experimental effects with random effects, leading to biased estimates of Type I error rates

and statistical power (Barr et al., 2013; Judd et al., 2017). Linear mixed-effects models (LMEMs) avoid information loss caused by averaging across stimuli within conditions (as in repeated-measures ANOVA) and can flexibly account for random effects arising from various sources (e.g., random sampling of stimuli, nested structures of participants). Consequently, LMEMs have become increasingly widespread in psychological experiments (Barr et al., 2013; Brauer & Curtin, 2018; Judd et al., 2017; Lee, 2018). A search of Web of Science for experimental psychology papers published in the last five years reveals that LMEMs are used approximately 1.5 times more frequently than ANOVA.

However, the application of LMEMs remains limited in China. For instance, among 181 experimental articles published in *Acta Psychologica Sinica*—China' s top psychology journal—between 2020 and 2022, only nine used LMEMs. Of these, five provided no justification for sample size determination, three used G*Power to approximate required sample size, and only one used the simr package with simulation-based power analysis to determine sample size. A major barrier to broader adoption is that the inclusion of random effects increases model complexity, rendering conventional sample size planning software (e.g., G*Power) inadequate. Researchers often feel uncertain about how to scientifically plan experimental designs and set appropriate numbers of participants and trials for LMEMs, creating a need for accessible programs or visualizations to guide sample size planning.

1.2 Sample Size Planning Based on Power Analysis

Traditional sample size planning primarily relies on power analysis for null hypothesis significance testing (NHST), requiring that sample size achieves a predetermined power level. Power analysis can be conducted through formula derivation or Monte Carlo simulation methods (e.g., Arend & Schäfer, 2019). Formula-based methods involve strong distributional assumptions and may yield biased results when data violate these assumptions (Judd et al., 2017). Monte Carlo simulation methods generate data repeatedly under specified parameters for a given model, fit the model to each simulated dataset, and calculate the proportion of replications yielding significant results. These methods do not require deriving parameter distributions, can handle non-normal data, and allow flexible model specification. Researchers have developed mature R packages (e.g., simr) that use Monte Carlo simulation to calculate power for LMEMs (Green & MacLeod, 2016).

To facilitate power-based sample size determination for nested data, researchers have developed intuitive visualizations and accompanying programs that display power across different sample sizes. The most widely used approach employs line graphs with sample size on the x-axis and power on the y-axis (e.g., Kumle et al., 2021). Researchers draw a horizontal line at the target power level, and the intersection with the power curve indicates the minimum required sample size. Murayama et al. (2022) developed an online program for generating such power curves. However, nested data requires determining sample sizes at two

levels, and the cost of increasing sample size differs across levels. Line graphs can only fix sample size at one level while varying the other, failing to simultaneously display the relationship between both level sample sizes and power. Schultzberg and Muthén (2018) used level-1 and level-2 sample sizes as x- and y-axes, respectively, with shaded regions indicating combinations meeting power criteria. Baker et al. (2021) proposed power contour plots that connect points with equal power across combinations of level-1 and level-2 sample sizes, using multiple contours to represent different power levels. In summary, for nested data, researchers need to observe the compensatory relationship between the two level sample sizes in terms of power and make cost-conscious trade-offs to determine appropriate sample sizes at each level.

1.3 Sample Size Planning Based on Effect Size Accuracy Analysis

The visualizations described above consider only statistical power. However, as criticism of NHST has grown, the American Statistical Association issued a statement on the cautious use of NHST, emphasizing that researchers should avoid reporting only significance and should also report effect sizes (Wasserstein & Lazar, 2016) and their interval estimates. Consequently, some scholars have proposed sample size planning based on effect size accuracy analysis.

The core of effect size accuracy analysis is controlling the width of the confidence interval (CI) for the effect size, with narrower intervals indicating more precise estimation (Maxwell et al., 2008). Some studies derive acceptable maximum CI width based on desired CI limits (Usami, 2020). For example, if a point estimate of effect size is 0.5 and its 95% CI width is 0.6, the 95% CI would be approximately [0.2, 0.8]. According to Cohen' s (2013) benchmarks, this interval covers small, medium, and large effect sizes (0.2, 0.5, 0.8), indicating poor precision (Maxwell et al., 2008; Usami, 2020). Other studies calculate minimum sample sizes directly from different CI widths (e.g., Kelley & Rausch, 2006). However, consensus has not yet been reached on how to determine an acceptable maximum CI width (e.g., Kelley et al., 2018).

To facilitate effect size accuracy-based sample size planning for nested data, Hecht and Zitzmann (2021) proposed overall performance plots using numbers of participants and time points as x- and y-axes, respectively. They calculated composite performance scores based on convergence rates, parameter estimation bias, and other indicators, using color blocks to differentiate scores. Researchers can weigh trade-offs based on these color blocks to determine appropriate sample size combinations. However, these plots do not consider statistical power, and the color blocks represent composite scores with some subjectivity, preventing researchers from clearly understanding the accuracy of parameters of interest.

1.4 Research Questions

In summary, sample size planning for nested data must simultaneously ensure adequate statistical power and effect size accuracy. However, existing methods,

programs, and visualizations typically address only one of these objectives (e.g., Arend & Schäfer, 2019; Kumle et al., 2021; Usami, 2020). No existing visualization enables researchers to simultaneously consider both requirements when planning sample size. Therefore, this study proposes CI width contour plots that employ Monte Carlo simulation for power and effect size accuracy analysis, simultaneously presenting both level sample size combinations with their corresponding power and CI width information. Since no universal standard exists for CI width, this study adopts two approaches from previous research, providing sample sizes for different CI widths and recommending that researchers derive acceptable maximum CI width from desired CI limits, then integrate power analysis results to identify optimal combinations of participants and trials.

Furthermore, in sample size planning for psychological experiments, researchers typically focus on fixed effects of experimental manipulations (Lee, 2018) while neglecting sample size planning for moderator effects of subject variables. However, as individual differences research has advanced, more studies explore whether experimental effects differ across individuals. For example, Jiang et al. (2022) found that participants' intertemporal decision-making (experimental effect) differed significantly between positive and negative emotional states (subject moderator variable). Such research requires sample size planning that ensures accurate estimation of subject variable moderation effects. Therefore, this study uses typical within-subject repeated experimental designs as a context to examine sample size planning based on LMEMs for both within-subject experimental effects and between-subject moderation effects.

This paper first reconstructs models within the multilevel modeling framework to better accommodate the need to add independent variables (control variables) at different levels. It then describes the procedure and functions for generating CI width contour plots. Finally, simulation studies based on within-subject experimental effects and between-subject moderation effects examine how experimental effects, random slopes, and subject variable types influence evaluation metrics and CI width contour plots, and demonstrate how to recommend appropriate sample sizes based on the results.

2 Linear Mixed-Effects Models in Psychological Experimental Research

The general form of LMEMs can be found in Williams et al. (2021). Within the multilevel modeling framework, they can be redefined. Consider a within-subject experimental design where stimuli are nested within experimental conditions and stimuli are not repeated (Barr et al., 2013; Lee, 2018). Level 1 represents the trial level, and level 2 represents the subject level, with trials nested within subjects. A random slope model (Model 1) can be expressed as:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + u_{0i} + u_{1i} X_{ij} + e_{0j} + \epsilon_{ij}$$

where Y_{ij} is the continuous outcome variable ($j = 1, \ldots, J$ denotes trials, $i = 1, \ldots, I$ denotes subjects), X_{ij} is the effect-coded experimental manipulation, u_{0i} and u_{1i} represent subject random intercepts and random slopes (capturing individual differences in baseline levels and experimental effects), and e_{0j} represents stimulus random intercepts (different stimuli have different effects). β_0 and β_1 are the means of subject random intercepts and random slopes, respectively, where β_1 is the fixed component and the primary effect size of interest. u_{0i}, u_{1i}, e_{0j}, and ϵ_{ij} represent random components at level 2 (intercept and slope), level 1 (stimulus), and residual errors. The model assumes $u_{0i} \sim N(0, \tau_{00})$, $u_{1i} \sim N(0, \tau_{11})$, $e_{0j} \sim N(0, \omega_{00})$, and $\epsilon_{ij} \sim N(0, \sigma^2)$.

A key advantage of multilevel models is the ability to conveniently add explanatory variables at different levels. For example, a subject variable Z_i can be added at level 2 to explain individual differences in random intercepts and random slopes (Model 2):

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \beta_{01} Z_i + \beta_{11} Z_{iX_{ij}} + u_{0i} + u_{1i} X_{ij} + e_{0j} + \epsilon_{ij}$$

where Z_i is a subject variable, β_{01} represents the effect of the subject variable on the random intercept, and β_{11} represents the effect of the subject variable on the random slope (also viewed as a cross-level interaction between level-1 and level-2 variables), which is the primary effect of interest. β_{11} represents the moderation effect of the subject variable on the experimental effect.

3 Procedure for Generating Confidence Interval Width Contour Plots

Sample size planning using simulation-based CI width contour plots involves the following steps:

First, set parameters. In an experimental research context, select a specific LMEM and specify level-1 sample size, level-2 sample size, fixed effect values, and random effect distributions.

Second, generate data. Based on the model defined in step one, repeatedly generate data N times (e.g., $N = 1000$).

Third, estimate parameters. For each replication, fit the model to the generated data. Use the R package lme4 (Bates et al., 2011) to estimate parameters via restricted maximum likelihood (REML). Calculate CIs for effect size parameters using the default Wald method.

Fourth, vary level-1 and level-2 sample sizes and repeat steps one through three.

Fifth, calculate evaluation metrics (see Section 4.2).

Sixth, evaluate metrics against criteria, draw CI width contour plots, and recommend appropriate sample sizes. This study suggests using the difference

between the highest and lowest effect size benchmarks as the acceptable maximum CI width.

This study developed an R function, `samplesize_{LMEM}.R`, for sample size planning in LMEMs (see online supplementary material 2). By calling this function with appropriate parameters, researchers can obtain evaluation metric results and CI width contour plots. The application workflow is shown in Figure 1 [Figure 1: see original paper]. Function call statements and explanations are provided in online supplementary material 3. The function offers flexibility: when $\omega_{00} = 0$, the data-generating model simplifies to one without stimulus random effects; when $\tau_{11} = 0$, it simplifies to a random intercept model; when all random effects are 0, it simplifies to a general linear model.

The following two simulation studies examine how different factors affect power and effect size estimation accuracy, illustrating the application of 95% CI width contour plots (this study uses 95% CIs) in sample size planning.

4 Simulation Study 1: Sample Size Planning for Within-Subject Experimental Effects

Study 1 uses Model 1 to examine the experimental effect β_1—the fixed effect of the level-1 independent variable—and investigates how effect size and random slope magnitude influence model estimation results, providing sample size recommendations through CI width contour plots.

4.1.1 Fixed Parameter Settings

Data were simulated based on Model 1. Following Arend and Schäfer (2019), the fixed effect of the random intercept was set to 0, and residual variance σ^2 was fixed at 1. Pilot studies found that intraclass correlation coefficient (ICC) magnitude had no significant impact on power or parameter estimation bias, so ICC was fixed at a moderate level of 0.3 (Arend & Schäfer, 2019). Given residual variance $\sigma^2 = 1$, the standardized random slope variance τ^*_{11} was fixed at a moderate level ($\tau^*_{11} = 0.09$) using the formula $\tau^*_{11} = \tau_{11}/\sigma^2$ (Arend & Schäfer, 2019). For simplicity, the covariance between random intercept and random slope was fixed at 0 ($\tau_{01} = 0$). Stimulus random effects were fixed at a small level of 0.2 (Cho et al., 2017). Finally, based on residual variance, the random slope variance for the data-generating model was determined.

The level-1 independent variable X_{ij} was set as a binary variable (e.g., control vs. experimental group) using deviation coding (-0.5 and 0.5) (Barr et al., 2013; Lee, 2018). Each condition was simulated 1000 times (e.g., Zhang, 2014).

4.1.2 Manipulated Parameter Settings

Following Arend and Schäfer (2019), experimental effect size (β_1) was set at three levels: 0.2 (small), 0.5 (medium), and 0.8 (large). Sample size planning was conducted for each condition.

Level-1 sample size (J, number of trials) included 10 levels: 10, 20, 30, 50, 70, 100, 150, 200, 250, 300. Level-2 sample size (I, number of subjects) included 9 levels: 10, 30, 50, 70, 100, 200, 400, 600, 800. This created $10 \times 9 = 90$ sample size combinations.

Additionally, research shows that unequal trial numbers across conditions (unbalanced designs) yield lower power at equivalent total sample sizes (Kumle et al., 2021). Therefore, to examine the impact of unbalanced designs on sample size planning, we added conditions with unequal sample sizes across the two categories of the independent variable at the medium effect size level, using a 1:4 ratio following Kumle et al. (2021).

After completing parameter settings, the `samplesize_{LMEM}.R` function was called to obtain results.

4.2 Evaluation Metrics

Evaluation metrics included five aspects: (1) **Convergence rate**: the proportion of replications where parameter estimation converged, assessed using lme4's default Hessian test (Bates et al., 2011). All subsequent metrics were calculated based on converged replications only. (2) **Statistical power**: the proportion of converged replications where the 95% CI for β_1 excluded 0, with a criterion of ≥ 0.8. (3) **Effect size (fixed effect) estimation accuracy**: including bias, relative parameter estimation bias (rbias), root mean squared error (RMSE), CI width, and CI coverage probability (CP). For the nth replication, let $\hat{\beta}_{1n}$ be the estimate and c_n the convergence indicator ($c_n = 0$ for non-convergence, $c_n = 1$ for convergence). Bias, rbias, RMSE, width, and CP were calculated as:

$$\text{bias} = \frac{\sum_{n=1}^{N} c_n(\hat{\beta}_{1n} - \beta_1)}{\sum_{n=1}^{N} c_n}$$

$$\text{rbias} = \frac{\text{bias}}{\beta_1}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{n=1}^{N} c_n(\hat{\beta}_{1n} - \beta_1)^2}{\sum_{n=1}^{N} c_n}}$$

$$\text{width} = \frac{\sum_{n=1}^{N} c_n \times w_n}{\sum_{n=1}^{N} c_n}$$

where w_n is the CI width for the nth replication.

$$\text{CP} = \frac{\sum_{n=1}^{N} c_n \times I_n}{\sum_{n=1}^{N} c_n}$$

where I_n indicates whether the CI covers the true value ($I_n = 0$ for no coverage, $I_n = 1$ for coverage). If effect size estimation is accurate, bias should be near 0, rbias should be below the critical value of 0.1 (Koch et al., 2014), RMSE should be small, width should be narrow, and CP should be between 0.925 and 0.975 (Bradley, 1978). (4) **Standard error estimation accuracy**: calculated as the bias of the estimated standard error relative to the standard deviation of estimates (SE-SD bias). If the estimated standard error is accurate, SE-SD bias should be close to 0 (Schultzberg & Muthén, 2018). (5) **Random effect estimation accuracy**: rbias for variance estimates of random effects (τ_{00}, τ_{11}, ω_{00}), calculated similarly to formula (10).

4.3.1 Convergence

Supplementary Tables 1 and 2 (online supplementary material 1) present convergence rates for the random slope model (Model 1) under balanced and unbalanced designs. Convergence was generally not problematic, with rates above 0.7 in all conditions and above 0.9 when both level sample sizes were below 200. Effect size and balanced vs. unbalanced design had minimal impact on convergence rates.

4.3.2 Power Results

Power results for balanced designs are shown in Table 1. Larger effect sizes yielded greater power, requiring smaller sample sizes to meet the 0.8 criterion. For example, with a moderate number of subjects (200), an effect size of 0.2 required 200 trials to achieve power ≥ 0.8, whereas an effect size of 0.8 required only 20 trials. Unbalanced design results are in Supplementary Table 3 (online supplementary material 1). Unbalanced designs showed consistently lower power than balanced designs. For instance, with 10 subjects and target power of 0.8, balanced designs required 50 trials while unbalanced designs required 100 trials.

Table 1 Power for level-1 independent variable effect in linear mixed-effects model under balanced design conditions in Study 1

Note: J = level-1 sample size, I = level-2 sample size, ES = effect size of level-1 independent variable. Bolded values indicate power $\geq$ 0.8.

4.3.3 Effect Size and Standard Error Estimation Accuracy

Effect size magnitude did not significantly affect estimation accuracy. Table 2 presents effect size and standard error estimation accuracy results for the balanced design with medium effect size (0.5), showing only rbias, width, and SE-SD bias (other metrics in Supplementary Table 4 ; results for effect sizes 0.2 and 0.8 in Supplementary Tables 5 and 6, online supplementary material 1). All conditions showed rbias < 0.1. Supplementary Table 4 shows bias near 0, RMSE small (mostly < 0.3) and decreasing with larger level-1 and especially level-2 sample sizes, and coverage > 0.925 except when level-1 sample size was 10.

These results indicate accurate estimation of the level-1 independent variable' s fixed effect.

Using Cohen' s benchmarks of 0.2 and 0.8 for small and large effects, we defined acceptable maximum 95% CI width as 0.8 - 0.2 = 0.6. Table 3 shows that when level-1 sample size was 30 or below, 95% CI width exceeded 0.6, indicating large standard errors and wide CIs.

SE-SD bias fluctuated near 0 across conditions, indicating accurate standard error estimation. Supplementary Table 7 (online supplementary material 1) shows that unbalanced designs had larger RMSE and wider 95% CIs than balanced designs.

Table 2 Accuracy of fixed effect and standard error estimation for level-1 independent variable under balanced design with effect size = 0.5 in Study 1

Note: J = level-1 sample size, I = level-2 sample size. Bolded rbias values indicate < 0.1.

4.3.4 Random Effect Estimation Accuracy

Effect size magnitude minimally affected random effect estimation accuracy (Supplementary Tables 8 -11, online supplementary material 1). For balanced design with medium effect size (0.5), rbias for τ_{00} and τ_{11} estimates was < 0.1, though ω_{00} estimation was relatively less accurate. Supplementary Table 11 shows that unbalanced designs had larger estimation biases for τ_{00} and τ_{11} than balanced designs.

4.3.5 Sample Size Planning Recommendations

This study proposes using CI width contour plots for sample size recommendations. Effect size accuracy is primarily reflected through CI width. Since random effect variances can also serve as effect size indicators (Hox et al., 2017), sample size planning can combine power, random effect estimation accuracy, and CI width.

For medium level-1 independent variable effect size (0.5), Figure 2 Figure 2: see original paper shows a power + CI width contour plot, where shaded regions indicate conditions meeting the power ≥ 0.8 criterion. Figure 2(b) shows a power + random effect accuracy + CI width contour plot, where shaded regions indicate conditions meeting both power ≥ 0.8 and rbias < 0.1 for all random effect estimates. Different colors correspond to different CI widths.

Figure 2 reveals three key patterns. First, for power alone or power + random effect accuracy, the two level sample sizes show compensatory effects. However, when level-1 (trials) sample size is too small (e.g., < 30), increasing level-2 (subjects) sample size cannot achieve adequate power or power + random effect accuracy. Second, 95% CI width is more strongly influenced by level-1 sample size; when level-1 sample size is small (e.g., 10), increasing level-2 sample size

does little to reduce 95% CI width. Third, compared to panel (a), panel (b)' s shaded region shifts upward and rightward, indicating that adding random effect accuracy requirements makes criteria more stringent. Contour plots for small, medium, and large effect sizes are in Supplementary Figures 1–3 (online supplementary material 1). As effect size increases, shaded regions shift downward, reducing the required level-1 sample size.

To apply CI width contour plots, first identify the range meeting criteria (power ≥ 0.8, or power ≥ 0.8 + rbias < 0.1 for random effects) using shaded regions. Then, within shaded regions, compare CI widths to the acceptable maximum to determine appropriate sample size combinations. For example, based on Figure 2, to meet power ≥ 0.8 with 95% CI width ≤ 0.6, we recommend level-1 sample size = 50 and level-2 sample size = 30. To meet both power ≥ 0.8 and rbias < 0.1 for all random effects with 95% CI width ≤ 0.6, we recommend level-1 sample size = 50 and level-2 sample size = 400.

Supplementary Figure 3 [Figure 3: see original paper] shows that unbalanced designs shift shaded regions upward, requiring larger level-1 sample sizes (at least 50) to meet power criteria.

(a) Power + CI width contour plot
(b) Power + random effect accuracy + CI width contour plot

Figure 2 CI width contour plots for medium level-1 independent variable effect size under balanced design in Study 1

Note: In panel (a), shaded regions indicate conditions meeting power ≥ 0.8. In panel (b), shaded regions indicate conditions meeting both power ≥ 0.8 and rbias < 0.1 for all random effect estimates. Different 95% CI widths are represented by different colored contours. As shown in the legend, eight contours are arranged sequentially from 0.3 to 1.0 in increments of 0.1. For example, the contour for 0.3 indicates that the region above the line has 95% CI width ≤ 0.3.

5 Simulation Study 2: Sample Size Planning for Subject Variable Moderation Effects

Study 2 uses Model 2 to examine the moderation effect of subject variables (β_{11}, cross-level interaction), investigating how random slope variance magnitude and subject variable type influence model estimation results and providing sample size recommendations through CI width contour plots.

5.1.1 Fixed Parameter Settings

Given that subject variables Z_i may be categorical (e.g., gender) or continuous (e.g., emotional arousal) in practice, Study 2 includes two scenarios: Scenario 1 with binary Z_i using deviation coding (-0.5 and 0.5), and Scenario 2 with continuous Z_i following a standard normal distribution.

Similar to Study 1, the fixed effect of the random intercept β_0 was set to 0. Study 2 focuses on the main effect of Z_i, fixed at a medium level: $\beta_{01} = 0.5$ (Scenario 1) and $\beta_{01} = 0.3$ (Scenario 2). For simplicity, following common practice in power analysis studies (e.g., Arend & Schäfer, 2019), β_1 was also fixed at a medium level: $\beta_1 = 0.5$ (Scenario 1) and $\beta_1 = 0.3$ (Scenario 2) (Cohen, 2013).

As in Study 1, residual variance was set to $\sigma^2 = 1$. In Scenario 1, with τ^*_{11} at three levels–0.01 (small), 0.09 (medium), and 0.25 (large) (Arend & Schäfer, 2019)–the standardized cross-level interaction effect was adjusted using formula (14) to obtain fixed effect parameters for the data-generating model (Arend & Schäfer, 2019).

In multilevel models, $\beta^*_{11} = \beta_{11} \times \sqrt{\tau^*_{11}}/\sigma$ is a partially standardized regression coefficient (standardizing only the dependent variable). When Z_i is categorical, β^*_{11} represents a fully standardized coefficient. When Z_i is continuous and already standardized, β^*_{11} is partially standardized. Therefore, in Scenario 1, β_{11} took values of 0.05, 0.15, and 0.25 under the three random slope variance levels. In Scenario 2, with τ^*_{11} fixed at medium level (0.09), $\beta_{11} = 0.09$.

β_{11} represents the moderation effect of the subject variable on the experimental effect. In Scenario 1, β_{11} represents the difference in experimental effects between the two Z_i categories. In Scenario 2, β_{11} indicates that subjects with higher/lower Z_i values show larger/smaller differences between experimental conditions. ICC was fixed at a moderate level, and stimulus random effects were fixed at 0.2. Each condition was simulated $N = 1000$ times.

5.2.2 Manipulated Parameter Settings

In Scenario 1, sample size planning was conducted with τ^*_{11} at 0.01, 0.09, and 0.25. To examine the impact of unbalanced designs, we added conditions with unequal sample sizes across Z_i categories (1:4 ratio). Sample size settings matched Study 1. The `samplesize_{LMEM}.R` function was called to obtain results.

5.2 Evaluation Metrics

Same as Study 1.

5.3.1 Convergence

Convergence rates for Study 2 are in Supplementary Tables 12 and 13 (online supplementary material 1). When Z_i was categorical, convergence rates fell below 0.7 in some conditions, and with $I = 800$ and $J = 250$ or 300, fewer than half of replications converged. This suggests that using random slope models when τ^*_{11} is small may cause convergence problems. Other conditions showed no major convergence issues, with rates generally above 0.7. Whether Z_i was categorical or continuous and whether the design was balanced had minimal impact on convergence rates.

5.3.2 Power Results

Power results across conditions are in Supplementary Tables 14 and 15 (online supplementary material 1). Larger τ^*_{11} yielded greater power. Power was generally higher for continuous than categorical Z_i, likely because continuous variables provide more information. Power increased with larger sample sizes at both levels, especially level-2. Unlike Study 1, power in Study 2 was more strongly influenced by level-2 sample size because power was calculated for the level-2 independent variable, which depends more on subject number. Unbalanced designs showed lower power than balanced designs.

5.3.3 Effect Size and Standard Error Estimation Accuracy

Results for moderation effect size and standard error accuracy are in Supplementary Tables 16 -20 (online supplementary material 1). Across conditions, bias, rbias, 95% CP, and SE-SD bias were consistent and small. RMSE increased and 95% CI widened with larger τ^*_{11}.

Unlike Study 1, where level-1 independent variable estimation accuracy was more affected by level-1 sample size, cross-level interaction estimation accuracy in Study 2 was more affected by level-2 sample size. For categorical Z_i with $\tau^*_{11} = 0.09$, effect size benchmarks for small and large effects were 0.06 (0.2×0.3) and 0.24 (0.8×0.3) using formula (15). Acceptable maximum 95% CI width was defined as 0.24 - 0.06 = 0.18. Supplementary Table 17 shows that 95% CI width exceeded this threshold in many conditions, meeting the requirement only when level-2 sample size was 400 with level-1 sample size ≥ 50, or level-2 sample size was ≥ 600 with level-1 sample size ≥ 20.

For $\tau^*_{11} = 0.01$, small and large effect benchmarks were 0.02 (0.2×0.1) and 0.08 (0.8×0.1), yielding acceptable maximum widths of 0.06. For $\tau^*_{11} = 0.25$, benchmarks were 0.1 (0.2×0.5) and 0.4 (0.8×0.5), yielding acceptable maximum width of 0.3. More conditions met CI width requirements when τ^*_{11} was large than when it was small.

For continuous Z_i, bias, rbias, 95% CP, and SE-SD bias were consistent and small. RMSE was smaller (see Supplementary Table 18) and 95% CI narrower. Using formula (15), effect sizes for small and large effects were 0.03 (0.1×0.3) and 0.15 (0.5×0.3), with acceptable maximum 95% CI width of 0.12. Unbalanced designs showed larger RMSE and wider 95% CIs than balanced designs.

5.3.4 Random Effect Estimation Accuracy

Supplementary Tables 21 -25 (online supplementary material 1) present rbias results for random effect estimates. First, similar to Study 1, τ^*_{11} magnitude, variable type, and balanced vs. unbalanced design minimally affected residual variance estimation accuracy, with rbias < 0.1 across conditions. Second, for categorical Z_i, estimation accuracy for τ_{00} and τ_{11} improved with larger sample sizes. When τ^*_{11} was small, rbias for τ_{11} estimates exceeded 0.1 in nearly all

conditions, with bias calculations showing overestimation of τ_{11}. When τ^*_{11} was large, rbias for τ_{00} estimates exceeded 0.1 in all conditions, with bias calculations showing overestimation of τ_{00}. Third, for continuous Z_i, estimation accuracy for τ_{00} and τ_{11} was slightly better than for categorical Z_i.

5.3.5 Sample Size Planning Recommendations

Using balanced design with medium τ^*_{11} as an example, Figures 3 and 4 show CI width contour plots for categorical and continuous Z_i, respectively. Unlike Study 1, 95% CI width was more strongly influenced by level-2 sample size; when level-2 sample size was small, increasing level-1 sample size did little to reduce 95% CI width, likely because the effect of interest was a level-2 variable. Additionally, compared to Study 1, the shaded region meeting both power and random effect accuracy criteria (panel b) shifted only slightly upward relative to the region meeting only power criteria (panel a), suggesting these criteria were similarly stringent. However, shaded regions in Study 2 shifted further upward and rightward than in Study 1, indicating that larger sample size combinations were needed to meet requirements.

For both categorical and continuous Z_i, regions meeting power and power + random effect accuracy criteria were similar, though the region meeting power criteria shifted slightly downward for continuous Z_i, indicating slightly smaller required level-1 sample sizes. Continuous Z_i also produced narrower 95% CIs, shifting contours leftward.

Based on Figure 3, for categorical Z_i, to meet power ≥ 0.8 with 95% CI width ≤ 0.18, we recommend level-1 sample size = 50 and level-2 sample size = 400. To meet both power ≥ 0.8 and rbias < 0.1 for all random effects with 95% CI width ≤ 0.18, we recommend level-1 sample size = 50 and level-2 sample size = 400.

Based on Figure 4 [Figure 4: see original paper], for continuous Z_i, to meet power ≥ 0.8 with 95% CI width ≤ 0.12, we recommend level-1 sample size = 50 and level-2 sample size = 200. To meet both power ≥ 0.8 and rbias < 0.1 for all random effects with 95% CI width ≤ 0.12, we recommend either level-1 sample size = 100 and level-2 sample size = 200, or level-1 sample size = 50 and level-2 sample size = 400.

CI width contour plots for categorical Z_i with small and large τ^*_{11} are in Supplementary Figures 4 and 5 (online supplementary material 1). When τ^*_{11} was small, shaded regions shifted upward and rightward, requiring larger sample sizes. When τ^*_{11} was large, the region meeting power criteria shifted slightly downward, reducing required level-1 sample size, but no conditions simultaneously met both power ≥ 0.8 and rbias < 0.1 for all random effects.

Supplementary Figure 6 [Figure 6: see original paper] (online supplementary material 1) shows that unbalanced designs shift shaded regions rightward, requiring larger level-2 sample sizes (at least 400) to meet power criteria.

(a) Power + CI width contour plot
(b) Power + random effect accuracy + CI width contour plot

Figure 3 CI width contour plots for categorical Z_i under balanced design in Study 2

(a) Power + CI width contour plot
(b) Power + random effect accuracy + CI width contour plot

Figure 4 CI width contour plots for continuous Z_i under balanced design in Study 2

6 Example Demonstration

This section demonstrates how to use the developed function to generate CI width contour plots for sample size planning in practice.

Suppose a researcher wants to examine whether certain personality traits (e.g., honesty, morality, humor) influence attractiveness to opposite-sex individuals, referencing a similar study on loyalty and attractiveness (Xu et al., 2020). That study used a single-factor within-subject design with non-repeated stimuli, presenting participants with opposite-sex faces paired with sentences describing relationship loyalty, and asked them to rate attractiveness. Loyalty (loyal vs. disloyal) was a within-subject factor with 20 non-repeated stimuli per condition. Results showed significantly higher attractiveness ratings for loyal potential partners. Researchers can use our proposed method for sample size planning.

First, select parameters for data generation by drawing from similar published research. For Xu et al.' s (2020) original data, we fitted Model 1 with loyalty as the independent variable and standardized face attractiveness ratings as the dependent variable. Detailed code and results are in online supplementary material 4. Based on results, we obtained: $\beta_1 = 0.578$, $\beta_0 = 0$, $\tau_{00} = 0.223$, $\tau_{11} = 0.249$, $\omega_{00} = 0.779$, $\sigma^2 = 0.017$.

Second, set parameters and call the function to generate evaluation results and CI contour plots. Set replications to $N = 1000$, level-1 sample size to six levels: 40, 80, 120, 200, 300, 400, and level-2 sample size to six levels: 10, 30, 50, 70, 100, 200, with equal trial numbers across conditions. Acceptable maximum 95% CI width was 0.8 - 0.2 = 0.6. Preset contour levels for 95% CI width were `kd <- c(0.3,0.4,0.5,0.6,0.7,0.8)`. The function call is shown in Figure 5 [Figure 5: see original paper].

```
source("samplesize_{LMEM}.R")
N <- 1000
I <- c(10,30,50,70,100,200)
J <- c(40,80,120,200,300,400)
P1 <- 0.5
P2 <- 0.5
#input 95%CI breaks
```

```
kd <- c(0.3,0.4,0.5,0.6,0.7,0.8)
#Model1
getModelOne(I,J,P1,P2,N,0.5775,0,0.223098,0.24948,0.779,0.01706)
generatePicData("modelOne_{{{evaluation}}_{{accuracy}}}",kd,c(0, max(I)),c(0, max(J)),I,J,I)
```

Figure 5 Function call statements for sample size planning in example demonstration

Third, run the program to obtain the evaluation file "modelOne_{{{evaluation}}_{{accuracy}}}.csv" and the power + CI width contour plot (Figure 6). Based on the plot, to meet power $\geq$ 0.8 with 95% CI width $\leq$ 0.6, the minimum recommended sample sizes are: 80 trials with 20 subjects, 60 trials with 30 subjects, or 40 trials with 70 subjects.

Figure 6 Power + CI width contour plot for example demonstration

Note: Since no conditions simultaneously met both power $\geq$ 0.8 and rbias $<$ 0.1 for all random effects, a power + random effect accuracy + CI width contour plot could not be generated.

7 Discussion

This study addressed sample size planning for LMEMs using simulation methods, examining within-subject experimental effects and between-subject moderation effects through two simulation studies. We investigated how experimental effect size, random slope variance, subject variable type, and balanced vs. unbalanced designs influence sample size recommendations, demonstrating the application of CI width contour plots. The goal was to provide methodological guidance and practical tools for researchers. Key findings are summarized below.

First, regarding convergence, Model 1 showed virtually no convergence problems. For Model 2, some convergence issues occurred when random slope variance was small and a maximal model was fitted.

Second, regarding power, larger effect sizes yielded greater power. Power was lower for categorical than continuous subject variables. Balanced designs showed higher power than unbalanced designs. The relationship between power and sample size depended on the level of the examined effect: power for level-1 independent variables was primarily affected by level-1 sample size, while power for level-2 independent variables was more affected by level-2 sample size. The two level sample sizes showed compensatory effects, but increasing sample size at the level of the effect of interest better compensated for small sample sizes at the other level.

Third, regarding effect size and standard error estimation accuracy, fixed effects were accurately estimated across all conditions when the fitted model was correctly specified. However, CI width was affected by balanced vs. unbalanced

designs and random effects. Unbalanced designs produced wider CIs. For level-2 variable moderation effects, larger random slope variance produced wider CIs and larger standard errors. Standard error estimation accuracy was high across conditions.

Fourth, regarding random effect estimation accuracy, residual variance was accurately estimated. Random intercept and random slope variance estimation accuracy was affected by balanced vs. unbalanced designs and random slope variance magnitude. For models with only within-subject independent variables, unbalanced designs reduced estimation accuracy for random intercept and random slope variances. Larger random slope variance decreased random intercept variance estimation accuracy but increased random slope variance estimation accuracy. Small random slope variance led to overestimation of random slope variance, while large random slope variance led to overestimation of random intercept variance.

7.2 Practical Recommendations

This study aims to illustrate sample size planning methods using two typical LMEMs. Based on the research process and results, we offer the following recommendations.

First, sample size planning should integrate both power analysis and effect size accuracy analysis. Traditional sample size planning based solely on power analysis (e.g., Schultzberg & Muthén, 2018) ensures adequate power (≥ 0.8). However, as more journals and institutions call for reporting effect sizes and CIs alongside significance, effect size estimation accuracy has become increasingly important (Maxwell et al., 2008). Power analysis and CI width-based planning are related yet distinct. Both relate to standard error: in fixed-effects models, CI is defined as $\widehat{\beta} \pm 1.96 \times \text{SE}$. In random-effects models, random effect variances contribute to standard error calculations, yielding larger standard errors and wider CIs. Smaller standard errors produce narrower CIs and more accurate effect size estimates. Assuming a non-zero true effect, narrower CIs are less likely to include 0, yielding higher power (Cohn & Becker, 2003). However, while larger true effect sizes increase power (as CIs are less likely to include 0), they do not affect CI width. Thus, larger effect sizes reduce sample sizes needed for adequate power but do not change sample size requirements based on CI width, consistent with our findings. This study found that sample size recommendations based on power analysis and effect size accuracy do not always coincide. For example, Figure 2(b) shows that for medium level-1 effect size, with level-2 sample size = 50, only 30 trials were needed for power > 0.8, but the 95% CI width was approximately 0.7, exceeding the acceptable maximum. Therefore, both criteria should be integrated when determining final sample size recommendations.

Second, when using simulation-based sample size planning, carefully determine parameters for the data-generating model. Power and effect size accuracy anal-

ysis require prespecifying model parameters (e.g., expected effect size, ICC) to generate data under a specific model. We emphasize that this study' s primary purpose is to illustrate methods and CI contour plot usage; our parameter settings may not represent most real-world situations. In practice, researchers can obtain these values from previously published similar studies, pilot data, meta-analyses, or expert opinions about minimally important effects (Pek & Park, 2019). However, some researchers note that using point estimates of effect size as true values ignores uncertainty regarding the unknown population effect size (Pek & Park, 2019), potentially yielding biased results. Therefore, some advocate methods that account for uncertainty, such as Bayesian hybrid approaches (Pek & Park, 2019).

Third, applied researchers can determine recommended sample sizes using our two types of CI width contour plots based on specific research needs. Building on Baker et al.' s (2021) power contour plots, we propose CI width contour plots that enable researchers to simultaneously consider multiple requirements and identify optimal sample sizes. Researchers can choose which plot to use based on their needs. If only power and effect size estimation accuracy are of interest, use power + CI width contour plots. If random effect estimation accuracy is also important—for example, to further analyze individual differences (e.g., using mixed-effects location-scale models, Williams et al., 2021) or to accurately calculate R^2 measures that incorporate random effects (e.g., Rights & Sterba, 2019)—use power + random effect accuracy + CI width contour plots. For CI width criteria, researchers can follow our approach, reference CI widths from previous studies, or determine critical values based on their study' s precision needs.

Finally, in practice, sample size planning involves comprehensive consideration of power, effect size accuracy, and research costs. Considering only power and effect size accuracy often leads to large recommended sample sizes, substantially increasing costs. This is especially impractical for resource-intensive studies (e.g., fMRI research). Therefore, some researchers have proposed methods that incorporate cost functions to identify sample sizes that meet power requirements while minimizing cost (e.g., Baker et al., 2021). For example, Baker et al.' s (2021) web application incorporates per-subject costs to calculate recommended sample sizes that achieve 80% power at minimum cost. Beyond cost, real-world sample size determination involves various constraints and prioritized requirements. Applied researchers can adapt our methods to their specific needs.

7.3 Future Directions

This study has limitations that future research could address. First, our simulation studies examined only experimental effect size, random slope magnitude, subject variable type, and balanced vs. unbalanced designs, fixing many other factors. Future research could investigate effects of covariance between random intercepts and slopes, stimulus random effect variance, and other factors to enrich findings. Second, this study focused on within-subject designs where

stimuli are nested within experimental conditions without stimulus-by-condition interactions, assuming binary experimental conditions and continuous outcomes. Future research could extend to other designs, continuous independent variables, categorical outcomes, and other scenarios to expand function capabilities. Finally, this study did not address uncertainty in expected effect sizes, failing to reflect real-world design challenges. Future research could adopt Pek and Park' s (2019, 2023) approach of using power and effect size accuracy distributions for sample size planning.

References

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.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255−278.

Baker, D. H., Vilidaite, G., Lygo, F. A., Smith, A. K., Flack, T. R., Gouws, A. D., & Andrews, T. J. (2021). Power contours: Optimising sample size and precision in experimental psychology and human neuroscience. *Psychological Methods, 26*(3), 295−314.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., ···& Grothendieck, G. (2011). Package 'lme4' . Linear mixed-effects models using S4 classes. R package version.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*, 144−152.

Brauer, M., & Curtin, J. J. (2018). Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychological Methods, 23*(3), 389−411.

Cho, S. J., De Boeck, P., & Lee, W. Y. (2017). Evaluating testing, profile likelihood confidence interval estimation, and model comparisons for item covariate effects in linear logistic test models. *Applied Psychological Measurement, 41*(5), 353−371.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.

Cohn, L. D., & Becker, B. J. (2003). How meta-analysis increases statistical power. *Psychological Methods, 8*(3), 243−253.

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.

Hecht, M., & Zitzmann, S. (2021). Sample size recommendations for continuous-time models: Compensating shorter time series with larger numbers of persons and vice versa. *Structural Equation Modeling: A Multidisciplinary Journal, 28*(2), 229–236.

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. Routledge.

Jiang, Y., Jiang, C., Hu, T., & Sun, H. (2022). Effects of emotion on intertemporal decision-making: Explanation from the single dimension priority model. *Acta Psychologica Sinica, 54*(2), 122−140. [蒋元萍, 江程铭, 胡天翊, 孙红月. (2022). 情绪对跨期决策的影响: 来自单维占优模型的解释. 心理学报, 54(2), 122−140.]

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: designs, analytic models, and statistical power. *Annual Review of Psychology, 68*(1), 601−625.

Kelley, K., Darku, F. B., & Chattopadhyay, B. (2018). Accuracy in parameter estimation for a general class of effect sizes: A sequential approach. *Psychological Methods, 23*(2), 226−243.

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.

Koch, T., Schultze, M., Eid, M., & Geiser, C. (2014). A longitudinal multilevel CFA-MTMM model for interchangeable and structurally different methods. *Frontiers in Psychology, 5*, 311. https://doi.org/10.3389/fpsyg.2014.00311

Kumle, L., Võ, 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.

Lee, W. Y. (2018). *Generalized linear mixed effect models with crossed random effects for experimental designs having non-repeated items: Model specification and selection* (Unpublished Doctoral dissertation). Vanderbilt University.

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*, 537−563.

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, 27*(6), 1014−1038.

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Almenberg, A. D., ···& Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology, 73*, 719−748.

Pek, J., & Park, J. (2019). Complexities in power analysis: Quantifying uncertainties with a Bayesian-classical hybrid approach. *Psychological Methods, 24*(5), 590−605.

Park, J., & Pek, J. (2023). Conducting Bayesian-classical hybrid power analysis with R package hybridpower. *Multivariate Behavioral Research, 58*(3), 543−559.

R Development Core Team. (2019). *R: A language and environment for statistical computing* [Computer software manual]. Vienna, Austria. http://www.Rproject.org (ISBN 3-900051-07-0).

Rights, J. D., & Sterba, S. K. (2019). Quantifying explained variance in multilevel models: An integrative framework for defining R-squared measures. *Psychological Methods, 24*(3), 309−338.

Schultzberg, M., & Muthén, B. (2018). Number of subjects and time points needed for multilevel time-series analysis: A simulation study of dynamic structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 25*(4), 495−515.

Usami, S. (2020). Confidence interval-based sample size determination formulas and some mathematical properties for hierarchical data. *British Journal of Mathematical and Statistical Psychology, 73*, 1−31.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: context, process, and purpose. *The American Statistician, 70*(2), 129-133.

Williams, D. R., Mulder, J., Rouder, J. N., & Rast, P. (2021). Beneath the surface: Unearthing within-person variability and mean relations with Bayesian mixed models. *Psychological Methods, 26*(1), 74−89.

Xu, L., Becker, B., Luo, R., Zheng, X., Zhao, W., Zhang, Q., & Kendrick, K. M. (2020). Oxytocin amplifies sex differences in human mate choice. *Psychoneuroendocrinology, 112*, 104483. https://doi.org/10.1016/j.psyneuen.2019.104483

Zhang, Z. (2014). Monte Carlo based statistical power analysis for mediation models: Methods and software. *Behavior Research Methods, 46*(4), 1184−1198.

Note: Figure translations are in progress. See original paper for figures.