Formulation of Mixed-Effects Mean-Variance Models and Exploration of Sample Size Planning

Authors: Liu Yue, Fang Fan, Liu Hongyun, Lei Yi, Liu Hongyun, Lei Yi

Date: 2023-01-31T00:00:00+00:00

Abstract

As research questions deepen and data collection methods advance, Mixed-Effects Location-Scale Models (MELSM), capable of appropriately analyzing and thoroughly extracting information from nested-structured data, have attracted widespread attention. This study aims to investigate MELSM model construction methods within a Bayesian framework through simulation studies and applied research, and to explore sample size planning paradigms for MELSM that integrate statistical power and effect size accuracy analysis under both certain and uncertain conditions. Ultimately, by integrating these functionalities, we will develop user-friendly software packages, establish an application workflow for MELSM, promote the adoption of new methods and technologies in psychological research, enhance the ecological validity and replicability of research, and thereby improve the overall quality of research.

Full Text

Model Construction and Sample Size Planning for Mixed-Effects Location-Scale Models

LIU Yue[1], FANG Fan[1], LIU Hongyun[2],[3], LEI Yi[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

Abstract

With increasing sophistication in research questions and advances in data collection methods, Mixed-Effects Location-Scale Models (MELSM) have garnered

widespread attention for their ability to appropriately analyze and deeply extract information from nested data structures. This study investigates MELSM model construction methods within a Bayesian framework through simulation and applied research, and explores sample size planning paradigms for MELSM that integrate statistical power and effect size accuracy analyses under both certain and uncertain conditions. Ultimately, we will develop a user-friendly software package that integrates these functionalities, establishing a comprehensive application workflow for MELSM. This will facilitate the adoption of new methods and techniques in psychological research, enhance ecological validity and replicability, and improve overall research quality.

Keywords: nested data, mixed-effects location-scale models, model construction, sample size planning

1. Research Background

In psychological and educational research, data often exhibit hierarchical nested structures. For example, in repeated measures designs, trials are nested within subjects; in longitudinal studies, measurement time points are nested within individuals; and in educational research, students are nested within classrooms. Such multi-level structured data are referred to as nested data. Nested data pose challenges for traditional data analysis methods. First, due to the non-independence of observations within the same group, nested data violate the assumptions of traditional statistical methods such as t-tests, ANOVA, and regression analysis, resulting in biased findings. Second, to satisfy experimental control and causal inference requirements, researchers need to include control or predictor variables at different levels (e.g., trial level, subject level), which traditional methods struggle to accommodate. Consequently, an increasing number of researchers recommend using Linear Mixed-Effects Models (LMM) for analyzing nested data (Hox et al., 2017). However, the homogeneity of residual variance assumption (hereinafter referred to as "residual homogeneity") in these models is frequently violated in practice. For instance, Williams et al. (2021) found significant individual differences in response time variability during conflict tasks examining cognitive control (e.g., Stroop task, Flanker task). Ignoring residual heterogeneity not only introduces parameter estimation bias but also hinders researchers from extracting valuable information about the stability of psychological traits (Williams et al., 2021). In summary, analytical methods for nested data should not only examine trait development trends and their influencing factors (between individuals) but also investigate trait stability during development and its influencing factors (within individuals), thereby providing rich evidence for revealing the nature of psychological phenomena. This is not only unattainable with traditional statistical methods but also presents new challenges for linear mixed-effects models.

To avoid biased estimation caused by residual heterogeneity and to flexibly explore mutual influences among behavioral traits and between/within-individual differences and their determinants, researchers have extended linear mixed-

effects models to propose a series of generalized models collectively known as Mixed-Effects Location-Scale Models (MELSM). MELSM does not require the residual homogeneity assumption and can explain interactions among traits at different levels, examine factors influencing trait variability and stability, and thus fully account for the impact of nested structures to yield richer research findings (Williams et al., 2021; Williams et al., 2019). Nevertheless, researchers face difficulties when applying MELSM. On one hand, since MELSM allows researchers to consider more random effects, omitting necessary random effects can increase Type I error rates (Barr et al., 2013), while including unnecessary random effects makes the model overly complex, causing parameter estimation difficulties and reducing statistical power (Judd et al., 2017; Lee, 2018). Therefore, how should researchers determine which random effects to include to construct an appropriate model? On the other hand, existing sample size planning procedures (e.g., G*Power, Faul et al., 2009; Faul et al., 2007) cannot be applied to MELSM. How then should researchers determine appropriate sample sizes for MELSM to ensure the replicability of findings and generalizability of conclusions (Nosek et al., 2022)? In summary, addressing the issues of model construction and sample size planning for MELSM is the primary task for promoting its application in psychological research.

2.1. Origin and Application of Mixed-Effects Location-Scale Models

To address the problem of residual variance heterogeneity in traditional linear mixed-effects models, researchers proposed the more generalized MELSM. This model consists of two components: the Location Model (describing the mean, referred to as the mean model based on its meaning) and the Scale Model (describing the scale, referred to as the variance model based on its meaning).

The location model corresponds to the mean component of linear mixed-effects models. Using the example of measurements (Level 1) nested within individuals (Level 2), its general form can be expressed as (Williams et al., 2021):

$$Y_i = X_i\beta + Z_i b_i + e_i$$

where Y_i is an $n_i \times 1$ column vector representing the outcome variable for individual i at Level 2, with n_i denoting the number of Level 1 measurements for individual i; X_i is an $n_i \times p$ matrix whose first column is 1 (representing the intercept for individual i) and columns 2 through p contain $p-1$ predictor variables; β is a $p \times 1$ column vector representing the fixed effects of the intercept and $p-1$ predictors; Z_i is an $n_i \times q$ matrix whose first column is 1 (representing the intercept for individual i) and columns 2 through q contain $q-1$ Level 1 predictors with random effects; b_i is a $q \times 1$ column vector representing the random effects of the intercept and $q-1$ predictors; and e_i is an $n_i \times 1$ column vector of residuals, $e_i \sim N_{n_i}(0, R_i)$, where R_i is an $n_i \times n_i$ covariance matrix. $R_i = \sigma_e^2 I_{n_i}$, where I_{n_i} is an $n_i \times n_i$ identity matrix. Linear mixed-effects models

typically assume that R_i satisfies the homogeneity assumption, meaning residuals are conditionally independent given the random effects. However, numerous studies have found this assumption is often violated in practice (Hedeker et al., 2008). Therefore, MELSM relaxes the residual homogeneity constraint by allowing heterogeneity of Level 1 residual variance in the scale model.

The scale model is defined as follows:

$$\sigma_{ei}^2 = \exp(W_i\tau + A_i t_i)$$

where σ_{ei}^2 is an $n_i \times 1$ column vector representing the residual variance for individual i; W_i is an $n_i \times s$ matrix whose first column is 1 (representing the intercept for individual i) and columns 2 through s contain $s-1$ predictor variables; τ is an $s \times 1$ column vector representing the fixed effects of the intercept and $s-1$ predictors; A_i is an $n_i \times a$ matrix whose first column is 1 (representing the intercept for individual i) and columns 2 through a contain $a-1$ Level 1 predictors with random effects; and t_i is an $a \times 1$ column vector representing the random effects of the intercept and $a-1$ predictors. The scale model effectively avoids bias in parameter estimation caused by residual heterogeneity and explains the sources of heterogeneity. For example, when applied to experimental research with trials nested within subjects, this model can explore why some individuals exhibit greater (or lesser) within-person response variability. Furthermore, one can add predictors to the variance of Level 2 random effects in the location model to explore factors influencing between-individual differences (Blozis et al., 2020). When applied to developmental and educational research with students nested within schools, this model can simultaneously explore factors affecting achievement variability both within and between schools, helping to improve instruction and promote educational equity (Williams et al., 2022).

Within the overall MELSM framework, researchers can also compute correlations between random effects in the location and scale models to further enrich research findings. For instance, in experimental studies with trials nested within subjects, the correlation between the random slope in the location model and the random intercept in the scale model describes whether subjects with stronger experimental effects tend to have more consistent (or inconsistent) responses. Additionally, one can include predictors to explain the covariances among random effects.

MELSM exhibits strong extensibility. Many researchers have developed rich extensions of MELSM tailored to different research questions. For example, nonlinear components can be added to both the location and scale models to reflect individual differences in learning trajectories and explore their influencing factors (Williams et al., 2019). Moreover, extensions have been developed for ordinal outcome variables (Hedeker et al., 2016), time-to-event censored data (Courvoisier et al., 2019), semi-continuous variables (e.g., data with many zeros, Blozis et al., 2020), as well as for dynamic data (Rast & Ferrer, 2018), cross-classified data (Brunton-Smith et al., 2017), and three-level nested designs (Lin

et al., 2018). Overall, the advantage of MELSM lies in its ability to simultaneously examine factors influencing both the development of dependent variables and the consistency of their variability, making it increasingly popular in psychological experimental research, longitudinal studies, and other fields.

2.2. Current Status of Mixed-Effects Location-Scale Model Construction

Both the location and scale models in MELSM may contain random effects, and selecting appropriate random effects for model construction represents the primary challenge in applying these models.

Misspecifying the model can lead to erroneous conclusions. On one hand, omitting necessary random effects produces biased results. Barr et al. (2013) recommended using models with as many random effects as possible when no theoretical hypotheses exist. González et al. (2014) supported this view, finding that omitting necessary random effects violates assumptions of residual independence, normality, and homogeneity, ultimately leading to incorrect standard error estimates. On the other hand, incorrectly specifying random effects also has serious consequences. Lee (2018) argued that the random structure of linear mixed-effects models should be correctly specified. Treating fixed effects as random effects increases parameter estimation error, can produce negative variance estimates (Baird & Maxwell, 2016), and reduces model power (Matuschek et al., 2017). Including too many random effects can cause model non-convergence (Judd et al., 2012). Therefore, both omitting existing random effects and including unnecessary ones can adversely affect parameter estimation, and appropriate random effects should be determined through model selection to construct the correct model (Brauer & Curtin, 2018; Martínez-Huertas et al., 2021).

Model construction should first consider theoretical hypotheses about random effects based on research design. When sufficient theoretical justification is lacking, data-driven approaches should be considered for model selection. Currently, nearly all existing MELSM studies directly specify models, and even when model selection is performed, they only examine whether heteroscedasticity exists in the scale model (Williams et al., 2021) or whether nonlinear components and their corresponding random effects are warranted (Williams et al., 2019). No studies have investigated model selection methods within a typical, complete MELSM framework.

Furthermore, exploring model selection and construction methods requires appropriate parameter estimation methods for MELSM. Within the maximum likelihood estimation framework, likelihood ratio tests (LRT) and information criteria (e.g., AIC, BIC) can be used for model comparison and selection in linear mixed-effects models (Lee, 2018). However, the high complexity of MELSM often leads to convergence difficulties with maximum likelihood estimation. Most existing MELSM studies have employed Bayesian methods for parameter estimation (e.g., Rast & Ferrer, 2018; Williams et al., 2020). Bayesian estimation

can flexibly handle complex models, but many conventional model selection methods are no longer applicable in this framework and require extension of traditional fit indices.

Bayesian fit indices can be divided into two categories. The first includes indices defined directly within the Bayesian framework. For example, the DIC (Deviance Information Criterion, Spiegelhalter et al., 2002) utilizes posterior distribution information to calculate model-data fit while including a penalty for model complexity. The Bayes factor improves upon p-value limitations in frequentist approaches by reflecting changes in updating prior probabilities to posterior probabilities given current data (Hoijtink et al., 2019). The Posterior Predictive p-value (PPP, Gelman et al., 1996) reflects the proportion of MCMC iterations where the posterior predictive discrepancy statistic exceeds the observed data discrepancy statistic. The second category extends approximate fit indices from structural equation modeling to the Bayesian framework. These approximate fit indices avoid strict testing deficiencies by tolerating small errors. Asparouhov and Muthén (2021) proposed methods for extending CFI, TLI, and RMSEA to the Bayesian framework, with the advantage of obtaining credible intervals for indices and using these intervals rather than point estimates for model comparison. Currently, no research has examined the performance of different Bayesian fit indices for MELSM model selection. This study aims to compare model selection results for MELSM using DIC, PPP, Bayes factors, CFI, TLI, and RMSEA.

2.3. Current Status of Sample Size Planning

The problem of insufficient statistical power due to small sample sizes is widespread across experimental research in various disciplines (Brysbaert & Stevens, 2018). Low power leads to poor replicability of p-value-based results (Hu et al., 2016). Most existing research has conducted power analysis for linear mixed-effects models to plan sample sizes. Only Walters et al. (2018) have addressed power for MELSM, but they focused solely on the power to detect residual heterogeneity or predictors in the scale model without predictors in the location model, and without random slopes for predictors in the scale model. They did not examine the power for fixed effects in the complete MELSM framework to achieve sample size planning.

Meanwhile, the American Statistical Association emphasizes avoiding sole reliance on significance reporting and incorporating examination of parameter estimation accuracy (Accuracy in Parameter Estimation, AIPE, primarily referring to effect size parameters, hereinafter referred to as "effect size accuracy") (Wen et al., 2016; Wasserstein & Lazar, 2016). In summary, sample size planning should satisfy not only power requirements but also be based on effect size accuracy analysis. The core of effect size accuracy analysis is controlling the width of effect size confidence intervals, with narrower intervals indicating more precise estimation. However, no research has yet integrated power analysis and effect size accuracy analysis for sample size planning in MELSM.

Traditional sample size planning research often faces three sources of uncertainty (Pek & Park, 2019, 2022). (1) Uncertainty in population effect sizes. For example, to achieve sample size planning based on regression models, researchers often use point estimates of regression coefficients from pilot or previous studies as substitutes for the true regression coefficients (population effect sizes). However, different samples (studies) yield different regression coefficients, creating uncertainty about the population effect size. (2) Uncertainty due to sample variation. That is, the specific samples used in different studies vary, yet sample size planning does not consider sample characteristics and can only provide general recommendations. (3) Uncertainty due to model selection. When multiple candidate models exist (as in MELSM construction), researchers typically presuppose the selected model is correct for sample size planning. However, the model chosen during data analysis may not correctly represent the actual data structure, introducing uncertainty from model selection.

Common Monte Carlo simulation-based power analysis paradigms can only address uncertainty from sample variation through repeated sampling, while ignoring uncertainty from population effect sizes and model selection, leading to inaccurate results. For example, Liu and Wang (2019) demonstrated that studies conducting sample size planning without considering uncertainty face underpowered consequences. Therefore, incorporating uncertainty issues into power and effect size accuracy analyses can better represent the practical dilemmas faced in research design and ensure more accurate and reliable sample size planning results, leading increasing methodological researchers to focus on sample size planning under uncertainty (e.g., Anderson et al., 2017; Liu & Wang, 2019). Pek and Park (2019, 2022) proposed a Bayesian-classical hybrid approach and developed corresponding software packages, providing a feasible path for addressing uncertainty issues. However, their research did not target MELSM, only considered power analysis without examining effect size accuracy, and used model averaging to address model selection uncertainty, which is less widely applied in practice than model selection (e.g., Barr et al., 2013; Lee, 2018).

3. Research Questions

Based on the theoretical and empirical status of MELSM, current issues regarding model construction and sample size planning for MELSM remain inadequately resolved. Specifically, these issues manifest in the following aspects:

First, most existing studies directly specify models without discussing model construction methods within a complete MELSM framework. Compared to linear mixed-effects models, MELSM can include random effects in the scale model, and the increase in candidate models creates greater difficulties for model selection (Williams et al., 2019). In practice, researchers often need to determine the inclusion of random effects through model selection. So, for complete MELSM, what sequence should researchers follow for model construction? Additionally, previous researchers have mostly discussed linear mixed-effects model construction using fit indices from the maximum likelihood estimation framework. How-

ever, convergence issues caused by complex random effect models have prompted consideration of more appropriate parameter estimation methods. Under the Bayesian estimation framework that is more suitable for MELSM, how do various fit indices perform? Research on these questions will effectively resolve model construction issues in MELSM applications.

Second, existing research has many shortcomings in sample size planning. First, power analysis for MELSM has only examined the power to detect residual heterogeneity or predictors in the scale model based on simple models, making it difficult to generalize findings to the complete MELSM framework, and unable to simultaneously examine the power for fixed effects in both location and scale models. Second, previous sample size planning has primarily relied on power analysis, with few studies proposing paradigms that integrate both power and effect size accuracy analysis for sample size planning, and no research exploring scientific sample size planning paradigms that combine both under uncertain conditions. This leads to low replicability in practice, with power and effect size accuracy failing to meet expected levels. Moreover, the Bayesian-classical hybrid method proposed by Pek and Park (2019, 2022) uses model averaging, which is less practically applicable than model selection (e.g., Barr et al., 2013; Lee, 2018), making resulting sample size recommendations less useful. Therefore, how can sample size planning for MELSM be conducted based on effect size accuracy analysis in addition to power analysis? Further, how can sample size planning for MELSM be conducted under uncertainty? How can the more commonly used model selection approach help resolve uncertainty from model selection? Research on these questions will help improve sample size planning theory and assist researchers in obtaining more reliable sample size recommendations.

Finally, previously developed software packages for mixed-effects models have limited functionality, capable of only one function among parameter estimation and construction, power analysis, or effect size accuracy analysis. No software package can flexibly implement MELSM parameter estimation and model selection, conduct both power and effect size accuracy analyses for sample size planning with and without consideration of uncertainty, and perform other related functions. To promote the widespread application of MELSM, researchers need a fully functional, user-friendly application for sample size planning, model construction, and other functions.

4. Research Design

Based on the application needs of MELSM, this study systematically investigates model selection/construction and sample size planning methods for MELSM. The overall approach combines theoretical and applied research, with the specific research flowchart shown in Figure 1 [Figure 1: see original paper]. Studies 1, 2, and 3 primarily employ theoretical derivation and simulation methods. Following Williams et al. (2021), the simulation studies examine scenarios where both location and scale models contain at most one

predictor, and the predictors are identical (the developed software package will accommodate more commonly used models). Considering model complexity, six nested models are primarily examined. Table 1 presents the main characteristics of each model and their nested relationships. To simplify the research, correlations between random effects in location and scale models are temporarily not considered (Arend & Schäfer, 2019). Model 1 contains no random effects; Model 2 adds a random intercept for the location model to Model 1; Model 3 adds a random slope for the location model predictor to Model 2; Model 4 adds a random intercept for the scale model to Model 3; Model 5 adds fixed slopes for scale model predictors to Model 4; and Model 6 adds a random slope for the scale model predictor to Model 5. Therefore, Models 1–3 assume residual homogeneity, while Models 4–6 assume residual homogeneity is violated to varying degrees.

Study 4 explores the application of sample size planning and model construction methods based on psychological experimental research and survey research. This study conducts a psychological experimental study based on the Stroop paradigm and an educational psychology survey investigating how mathematics learning self-efficacy and teachers' use of cognitive activation instructional strategies affect mathematics achievement.

The statistical software used in this study primarily includes the R package brms for Bayesian estimation (Bürkner, 2017) and a self-developed software package. Specific plans for each study are as follows.

4.1. Study 1: Model Selection and Construction for Mixed-Effects Location-Scale Models

Study 1 explores fundamental approaches to MELSM model construction based on Bayesian methods and corresponding fit indices suitable for complex model parameter estimation. The main idea is to generate data from true models containing different random effects (Models 1-6), then use Bayesian fit indices to compare candidate models and identify the data-supported model. Finally, by evaluating the consistency between selected models and true models across all simulation replications, we will summarize the advantages, disadvantages, and applicable ranges of each fit index to select robust indices for Study 3. Since simultaneously considering location and scale models involves many candidate models, this study adopts a simplified approach of first selecting the most appropriate location model, then selecting the most appropriate scale model, and examines the feasibility of this strategy. Study 1 consists of two sub-studies.

Study 1-1 includes two scenarios examining location model comparison and selection results when data do and do not satisfy residual homogeneity. Simulation conditions include Level 1 sample size (10, 30, 70, 100, 300), Level 2 sample size (20, 50, 300, 800), effect size of location model predictors (0.2—small, 0.5—medium, 0.8—large), and variance of location model random slopes (0.01, 0.09, 0.25), creating 5$\times$4$\times$3$\times$3 = 180 combinations of simulation conditions (i.e.,

cells). For Level 1 sample size, n=10 represents conditions where Lee (2018) found no convergence issues using Laplace approximation, while n=300 represents the maximum number of time points tested in Schultzberg and Muthén' s (2018) sample size planning study for dynamic structural equation models. For Level 2 sample size, N=20 approximates the minimum number of subjects (16) used in similar experimental designs summarized by Lee (2018), while N=800 approximates the 1,000-subject condition in Lee' s (2018) simulation study, aiming to explore the effect of large sample conditions on effect size estimation accuracy and power improvement. Variation between minimum and maximum sample sizes references similar sample size planning studies (e.g., Schultzberg & Muthén, 2018). Effect sizes for location model predictors reference Cohen' s d small, medium, and large levels (Barr et al., 2013; Lee, 2018). Variance of location model random slopes references levels set in Arend and Schäfer (2019). For each combination, following most power analysis studies, 1,000 datasets are generated based on each data-generating model (e.g., Thoemmes et al., 2010; Zhang, 2014). Scenario 1 generates data from Models 1-3, while Scenario 2 generates data from three models with the scale model of Model 4 and location models of Models 1-3. Candidate models fitted in both scenarios are Models 1-3.

Bayesian estimation is applied for parameter estimation. Appropriate prior distributions are determined based on sensitivity analysis results. Following previous similar studies (Depaoli & Clifton, 2015; van Erp & Browne, 2021), two prior distributions for variance are compared: non-informative priors (inverse Gamma distribution) and robust priors (mixture inverse Gamma distribution). For regression coefficients, standard normal distributions are used following similar studies (Depaoli & Clifton, 2015; van Erp & Browne, 2021). After model fitting, the best model is selected based on fit indices, and the proportion of correct model selections by each index is calculated under each data-generating model to identify robust fit indices. The compared fit indices include DIC, PPP, Bayes factors, CFI, TLI, and RMSEA.

Study 1-2 examines scale model comparison and selection results when the location model has been correctly specified and residual homogeneity is violated. Simulation conditions are identical to Study 1-1, creating 180 combinations of simulation conditions. For each combination, 1,000 datasets are generated based on Models 4-6. Candidate fitted models are Models 3-6, with the same fit indices and analysis procedures as Study 1-1.

4.2. Study 2: Statistical Power and Effect Size Accuracy for Mixed-Effects Location-Scale Models

Study 2 achieves sample size planning based on both power analysis and effect size accuracy analysis to ensure sample sizes satisfy both requirements. The main approach uses the same model for data generation and fitting, calculates power across different sample size conditions using Monte Carlo simulation, and computes 95% credible intervals for effect sizes using posterior distribution-

based methods. Study 2 consists of two sub-studies using Model 3 (MELSM with residual homogeneity, i.e., linear mixed-effects model) and Model 6 (full MELSM) as data-generating and fitting models. Power and effect size accuracy analyses target fixed effects of predictors in the location model for Study 2-1 and fixed effects of predictors in both location and scale models for Study 2-2.

Study 2-1 examines power and effect size accuracy for MELSM with residual homogeneity. Simulation conditions include Level 1 sample size (10, 25, 50, 75, 100, 150, 200, 300), Level 2 sample size (20, 30, 50, 75, 100, 150, 200, 300, 800), and effect size of location model predictors (0.2—small, 0.5—medium, 0.8—large), creating 8$×9×$3 = 216 combinations of simulation conditions. For each combination, 10,000 datasets are generated based on Model 3. The number of replications follows sensitivity analysis results from power analysis studies (Pek & Park, 2019, 2022). Model 3 is fitted, and power, 95% credible interval width of effect size estimates, and coverage rates of 95% credible intervals for the true value are calculated for fixed effects of location model predictors. Finally, recommended sample sizes are identified that simultaneously achieve power above 0.8, narrow 95% credible interval width, and coverage rates between 92.5% and 97.5%.

Study 2-2 examines power and effect size accuracy for full MELSM. Simulation conditions add effect size of scale model predictors (0.2—small, 0.5—medium, 0.8—large) to those in Study 2-1, creating 8$×9×3×$3 = 648 combinations of simulation conditions. For each combination, 10,000 datasets are generated based on Model 6. Model 6 is fitted, and analyses of fixed effects for predictors in both location and scale models follow the same procedures and evaluation criteria as Study 2-1.

4.3. Study 3: Sample Size Planning for Mixed-Effects Location-Scale Models Under Uncertainty

Study 3 primarily explores sample size planning under uncertain conditions, providing references for accurately solving research design problems in more practical and widespread scientific practice. The main idea extends the sample size planning method to uncertain situations based on robust fit indices selected in Study 1 and the paradigm integrating power and effect size accuracy analysis from Study 2. This method is then used to explore recommended sample sizes for MELSM with residual homogeneity and full MELSM under conditions with different degrees of effect size uncertainty.

The extended method proceeds as follows: (1) Define prior distributions for effect size parameters. Based on existing research, determine the possible range of effect sizes, then estimate the standard deviation of effect sizes assuming the range is approximately 6 standard deviations under normality. This yields normal distribution parameters for effect size prior distributions (other prior distributions can also be explored). (2) Draw possible effect sizes. Extract S effect size values from the prior distribution defined in (1). (3) Generate samples.

Use the most complex candidate model as the data-generating model (Pek & Park, 2019), generate R samples of size N using each drawn effect size as the generating value, with other parameter settings referencing Study 2, yielding R×S samples. (4) Construct models. Fit the R×S datasets with candidate models of varying complexity, apply robust fit indices selected in Study 1 to choose appropriate models, and calculate average power, 95% credible interval width, and coverage rate across datasets generated from each effect size (R datasets). (5) Integrate results. Integrate power and effect size accuracy results across S effect sizes to obtain distributions of these indices for sample size N.

Following sensitivity analysis results from Pek and Park (2019), set S=1,000 and R=10,000. This process will be programmed in R and included in the software package developed for this project.

The simulation study follows the sub-study classification of Study 2, divided into two sub-studies. Study 3-1 investigates sample size planning for MELSM with residual homogeneity under uncertainty. Simulation conditions include effect size of location model predictors (0.2—small, 0.5—medium, 0.8—large), degree of uncertainty in population effect size for the location model (range of effect size distribution = 0.15, 1.50, 3.00), Level 1 sample size (10, 25, 50, 75, 100, 150, 200, 300), and Level 2 sample size (20, 30, 50, 75, 100, 150, 200, 300, 800), creating 3$×3×8×$9 = 648 combinations of simulation conditions. The range of effect size distribution references settings from Pek and Park' s (2019) simulation study. Following the basic method for power and effect size accuracy analysis under uncertainty, for each combination, data are generated from Model 3, and models are constructed by comparing candidate Models 1-3. Finally, distributions of power and effect size accuracy indices for sample size N are obtained. Researchers can determine recommended sample sizes based on different criteria, such as the 20th percentile of the power distribution exceeding 0.8, or the mean power exceeding 0.8. The corresponding sample size that meets the requirements is the recommended value.

Study 3-2 investigates sample size planning for full MELSM under uncertainty. To simplify the study, the effect size for location model predictors is fixed at medium, and overall effect size uncertainty is fixed at medium level. Simulation conditions include effect size of scale model predictors (i.e., η_1, 0.2—small, 0.5—medium, 0.8—large), degree of uncertainty in population effect size for the scale model (range of effect size distribution = 0.15, 1.50, 3.00), Level 1 sample size (10, 25, 50, 75, 100, 150, 200, 300), and Level 2 sample size (20, 30, 50, 75, 100, 150, 200, 300, 800), creating 3$×3×8×$9 = 648 combinations of simulation conditions. Following the basic method for power and effect size accuracy analysis under uncertainty, for each combination, data are generated from Model 6. Consistent with the strategy for constructing full MELSM in Study 1, the location model is first determined by comparing candidate Models 1-3, then the scale model is determined by comparing candidate Models 3-6. Recommended sample size determination follows Study 3-1.

Finally, this study will integrate results from the first three studies to develop

a user-friendly software package that enables applied researchers to implement MELSM sample size planning, model construction, and data analysis, promoting the application of new methods and techniques in psychological research.

4.4. Study 4: Applied Research on Mixed-Effects Location-Scale Models

Study 4 demonstrates the standard workflow for MELSM sample size planning, model construction, and result interpretation through two practical psychological research questions, validating the operability of conclusions from the first three studies in real applications. The research process for both cases is as follows: (1) Before data collection, following the sample size planning paradigm under uncertainty from Study 3, use the software package developed in this project to input multiple candidate models, prior distributions of effect sizes, and combinations of Level 1 and Level 2 sample sizes based on previous relevant research results. Then determine appropriate Level 1 and Level 2 sample sizes according to the software output of effect size accuracy and power indices under each condition. (2) Refine the research design and implement the study according to the determined sample sizes to collect data. (3) Construct an appropriate MELSM based on the data, estimate parameters, interpret results, and draw conclusions.

Case 1 is a psychological experimental study based on the Stroop paradigm, aiming to examine the effects of congruent and incongruent conditions on correct response times and on the stability of correct response times. Research design: The experimental task is a numerical Stroop task with one within-subject independent variable having two levels: congruent and incongruent conditions. In congruent conditions, the number of characters matches the displayed digit (e.g., 333). In incongruent conditions, the number of characters does not match the displayed digit (e.g., 44). The task requires subjects to count the number of characters, with the dependent variable being response time for correct trials. The collected data have a nested structure with trials nested within subjects. The candidate models considered in this study are similar to Models 1–6 in the simulation study, with the independent variable in both location and scale models being experimental condition (congruent/incongruent).

Case 2 is an educational psychology survey exploring how mathematics learning self-efficacy and teachers' use of cognitive activation instructional strategies affect mathematics achievement, aiming to examine the effects of student mathematics learning self-efficacy and teachers' use of cognitive activation strategies on mathematics achievement and on within-class achievement consistency. Research design: First, questionnaires on student mathematics self-efficacy and teachers' use of cognitive activation instructional strategies will be developed and validated through pilot testing. Second, using stratified sampling in a district in Sichuan Province, primary schools will be selected first, then one fourth-grade class from each sampled school will be randomly selected. Students in these classes will complete the mathematics learning self-efficacy questionnaire,

their mathematics teachers will complete the cognitive activation instructional strategies questionnaire, and students' mathematics achievement scores from the district-wide unified examination will be obtained. The collected data have a nested structure with students nested within classes. Student mathematics learning self-efficacy is a Level 1 predictor, and teachers' use of cognitive activation instructional strategies is a Level 2 predictor. The candidate models considered in this study include (to simplify, interaction between self-efficacy and cognitive activation is not considered): Model 1—location model predictors are self-efficacy and cognitive activation, no random effects, assuming variance homogeneity; Model 2—adds random intercept for location model to Model 1; Model 3—adds random slope for self-efficacy in location model to Model 2; Model 4—adds random intercept for scale model to Model 3; Model 5—adds fixed slopes for self-efficacy and cognitive activation in scale model to Model 4; Model 6—adds random slope for self-efficacy in scale model to Model 5. Since only self-efficacy is a Level 1 predictor, only this variable may have random slopes.

5. Theoretical Contributions and Innovations

MELSM, which is suitable for situations with residual heterogeneity and can reasonably and deeply extract information from nested data, has received considerable attention from foreign researchers in recent years. Some studies have conducted theoretical research such as power analysis based on MELSM (Walters et al., 2018; Williams et al., 2021; Williams et al., 2019), while others have applied MELSM in empirical research to obtain rich findings (Rast & Ferrer, 2018; Williams et al., 2020). However, theoretical research and practical application of MELSM are still in their infancy both domestically and internationally. Issues such as sample size planning and model construction in practical applications remain inadequately resolved, leaving researchers uncertain when applying MELSM. To promote the widespread application of MELSM in psychological research, this study will first explore Bayesian fit indices suitable for complete MELSM model selection and propose MELSM model construction methods (Study 1). Then it will investigate sample size planning methods for MELSM based on both power analysis and effect size accuracy analysis under certain conditions (Study 2), and further apply the robust Bayesian fit indices obtained in Study 1 to sample size planning under uncertain conditions (Study 3), ultimately forming a paradigm for MELSM sample size planning and model construction and developing an integrated, user-friendly software package. Finally, this study will verify simulation results through empirical research and demonstrate the application workflow for MELSM (Study 4).

The theoretical paradigm for MELSM sample size planning and model construction proposed in this study is shown in Table 2 . After determining the research topic and completing the experimental design, a standard experimental study typically includes processes such as sample size planning, data collection, data analysis, and result interpretation. On one hand, researchers should conduct sample size planning before data collection to ensure the sample size meets power

and effect size accuracy requirements. Sample size planning faces three sources of uncertainty (Pek & Park, 2019). Monte Carlo-based analysis paradigms can handle uncertainty from sample variation through repeated sampling. When researchers can obtain reliable effect size estimates from pilot studies, previous research, or meta-analysis, uncertainty in population effect sizes may be disregarded. When researchers can determine the fitted model based on theory, uncertainty from model selection may be disregarded. Therefore, depending on whether these two sources of uncertainty exist, researchers can adopt different paradigms for sample size planning to determine the required sample size for subsequent data collection.

On the other hand, after data collection, researchers should construct appropriate models. If a specific model can be determined based on theory, model selection is unnecessary and the model can be fitted directly for result analysis. If model selection uncertainty exists, data-driven methods should be used for model selection. Specifically, researchers should first determine the optimal location model, then determine the optimal scale model based on the Bayesian fit indices recommended in this study, thereby obtaining the optimal MELSM for data analysis.

The innovations of this study are mainly reflected in two aspects. First, methodological paradigm innovation. This study fully considers parameter estimation methods suitable for MELSM, 首次探讨贝叶斯估计框架下的拟合指标在 MELSM 模型选择中的表现及其影响因素, and innovatively proposes a model construction approach that sequentially selects location and scale models within the complete MELSM framework. This will provide more reliable fit indices for MELSM model construction and meet the application needs of MELSM. Additionally, this study incorporates both power analysis and effect size accuracy analysis into sample size planning, and improves the sample size planning paradigm under uncertain conditions based on Pek and Park (2019, 2022). It summarizes sample size planning paradigms for four scenarios (effect size certain/uncertain $\times$ model certain/uncertain), further improving MELSM sample size planning methods, enriching sample size planning theory, and providing more scientific and reliable methodological recommendations for practical sample size planning, thereby enhancing the replicability of experimental research.

Second, practical application innovation. This study develops a convenient software package tailored to the characteristics of psychological research to meet the application needs of MELSM sample size planning and model construction under the Bayesian framework, providing a software foundation for the popularization of MELSM. This has guiding significance and innovative value for scientifically conducting research design and data analysis, representing an innovative exploration in the practice of psychology.

In summary, this study deeply explores MELSM model construction and sample size planning methods, providing methodological support for scientifically conducting psychological research. The findings will further promote the application of MELSM in psychological research, offering new perspectives for deeply

extracting information from nested data and revealing the nature of complex psychological phenomena.

References

胡传鹏, 王非, 过继成思, 宋梦迪, 隋洁, 彭凯平. (2016). 心理学研究中的可重复性问题：从危机到契机. 心理科学进展, 24(9), 1504-1518.

温忠麟, 范息涛, 叶宝娟, 陈宇帅. (2016). 从效应量应有的性质看中介效应量的合理性. 心理学报, 48(4), 435-443.

Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-size planning for more accurate statistical power: A method adjusting sample effect sizes for publication bias and uncertainty. *Psychological Science, 28*(11), 1547-1562.

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.

Asparouhov, T., & Muthén, B. (2021). Advances in Bayesian model fit evaluation for structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 28*(1), 1-14.

Baird, R., & Maxwell, S. E. (2016). Performance of time-varying predictors in multilevel models under an assumption of fixed or random effects. *Psychological Methods, 21*(2), 175-191.

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.

Blozis, S. A., McTernan, M., Harring, J. R., & Zheng, Q. (2020). Two-part mixed-effects location scale models. *Behavior Research Methods, 52*, 1836-1847.

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.

Brunton-Smith, I., Sturgis, P., & Leckie, G. (2017). Detecting and understanding interviewer effects on survey data by using a cross-classified mixed effects location–scale model. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 180*(2), 551-568.

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), 1-28.

Courvoisier, D., Walls, T. A., Cheval, B., & Hedeker, D. (2019). A mixed-effects location scale model for time-to-event data: A smoking behavior application. *Addictive Behaviors, 94*, 42-49.

Depaoli, S., & Clifton, J. P. (2015). A Bayesian approach to multilevel structural equation modeling with continuous and dichotomous outcomes. *Structural Equation Modeling: A Multidisciplinary Journal, 22*(3), 327–351.

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.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191.

Gelman, A., Meng, X. L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6*(4), 733–760.

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

Halsey, L. G., Curran-Everett, D., Vowler, S. L., & Drummond, G. B. (2015). The fickle p value generates irreproducible results. *Nature Methods, 12*(3), 179–185.

Hedeker, D., Mermelstein, R. J., & Demirtas, H. (2008). An application of a mixed-effects location scale model for analysis of ecological momentary assessment (EMA) data. *Biometrics, 64*(2), 627–634.

Hedeker, D., Mermelstein, R. J., Demirtas, H., & Berbaum, M. L. (2016). A mixed-effects location-scale model for ordinal questionnaire data. *Health Services and Outcomes Research Methodology, 16*(3), 117–131.

Hoijtink, H., Mulder, J., van Lissa, C., & Gu, X. (2019). A tutorial on testing hypotheses using the Bayes factor. *Psychological Methods, 24*(5), 539–556.

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications* (3rd ed.). New York, Routledge.

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality & Social Psychology, 103*(1), 54–69.

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.

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 doctorial dissertation). Vanderbilt University.

Lin, X., Mermelstein, R. J., & Hedeker, D. (2018). A 3-level Bayesian mixed effects location scale model with an application to ecological momentary assessment data. *Statistics in Medicine, 37*(13), 2108–2119.

Liu, X., & Wang, L. (2019). Sample size planning for detecting mediation effects: A power analysis procedure considering uncertainty in effect size estimates. *Multivariate Behavioral Research, 54*(6), 822–839.

Martínez-Huertas, J. Á., Olmos, R., & Ferrer, E. (2021). Model selection and model averaging for mixed-effects models with crossed random effects for subjects and items. *Multivariate Behavioral Research, 59*(4), 390–412.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language, 94*, 305–315.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9*(2), 147–163.

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. (2022). Conducting Bayesian-classical hybrid power analysis with R package hybridpower. *Multivariate Behavioral Research.* https://doi.org/10.1080/00273171.2022.2038056

Rast, P., & Ferrer, E. (2018). A mixed-effects location scale model for dyadic interactions. *Multivariate Behavioral Research, 53*(5), 756–775.

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.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B (Statistical Methodology), 64*(4), 583–639.

Thoemmes, F., MacKinnon, D. P., & Reiser, M. R. (2010). Power analysis for complex mediational designs using Monte Carlo methods. *Structural Equation Modeling, 17*(3), 510—534.

van Erp, S., & Browne, W. J. (2021). Bayesian multilevel structural equation modeling: An investigation into robust prior distributions for the doubly latent categorical model. *Structural Equation Modeling: A Multidisciplinary Journal, 28*(6), 875-893.

Walters, R. W., Hoffman, L., & Templin, J. (2018). The power to detect and predict individual differences in intra-individual variability using the mixed-effects location-scale model. *Multivariate Behavioral Research, 53*(3), 362–377.

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., Martin, S. R., Liu, S., & Rast, P. (2020). Bayesian multivariate mixed-effects location scale modeling of longitudinal relations among affective traits, states, and physical activity. *European Journal of Psychological Assessment, 36*(6), 981–997.

Williams, D. R., Martin, S. R., & Rast, P. (2022). Putting the individual into reliability: Bayesian testing of homogeneous within-person variance in hierarchical models. *Behavior Research Methods, 54*(3), 1272–1290.

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.

Williams, D. R., Zimprich, D. R., & Rast, P. (2019). A Bayesian nonlinear mixed-effects location scale model for learning. *Behavior Research Methods, 51*(5), 1968–1986.

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.