Extension and Application of Mixed-Effects Mean-Variance Models in Intervention Studies

Authors: Yueling He, Liu Yue, Liu Yue

Date: 2025-11-20T16:06:30+00:00

Abstract

Randomized controlled trial designs with pretest-posttest repeated measures, as the most scientifically rigorous research design, have been increasingly widely applied in intervention research. However, current data analysis methods fail to fully capture intervention effects: they neglect intra-individual variability (IIV), rely on overly simplistic evaluation metrics, and inadequately explore covariates. The lack of more advanced data analysis models to match this design greatly limits its practical adoption. Based on the mixed-effects location-scale model (MELSM), this study proposes a pretest-posttest MELSM specifically applicable to randomized controlled trial designs with pretest-posttest repeated measures (RCTRM). This paper first demonstrates the necessity of using this proposed model through empirical data, and then investigates its influencing factors using Monte Carlo simulation methods.

Full Text

Extension and Application of Mixed-Effects Location-Scale Models in Intervention Research

He Yueling[1], **Liu Yue**[1]* [1]Institute of Brain and Psychological Sciences, Sichuan Normal University, Chengdu 610066, China

The randomized controlled trial with pretest–posttest repeated measures (RCTRM) has increasingly become a widely used and scientifically rigorous design in intervention research. However, current analytic approaches have not fully captured intervention effects, as they often overlook intra-individual variability (IIV), rely on overly limited evaluation indicators, and insufficiently address covariates. The lack of advanced analytic models tailored to this design has substantially constrained its broader application in practice. To address these limitations, the present study proposes a mixed-effects location–scale model (MELSM) specifically adapted for RCTRM data, referred to as the pretest–

posttest MELSM. Using empirical data, we first illustrate the necessity of applying this model. We then conduct a Monte Carlo simulation study to investigate factors that influence the performance of the proposed model.

Keywords: Mixed-effects location-scale model Intensive longitudinal data Intra-individual variability RCT with repeated measures

In social and behavioral science research, intervention studies represent a widely adopted paradigm in psychology, clinical medicine, and related fields for addressing mental health problems (Agteren et al., 2021; Brailovskaia et al., 2020). By systematically manipulating independent variables (e.g., pharmacological interventions, psychotherapy, physical therapy) and observing changes in dependent variables (e.g., anxiety, depression), intervention studies provide critical evidence for understanding psychopathological mechanisms and developing intervention strategies (Kazdin, 2017).

1.1 Randomized Controlled Trials with Repeated Measures (RCTRM)

Intervention studies typically employ either non-randomized single-group designs or randomized controlled trials (RCTs). Non-randomized single-group designs are cost-effective (Shadish, Cook, & Campbell, 2002) but struggle to control for confounding factors such as time effects or selection bias, limiting their application in psychological research (Cook & Campbell, 1979). In contrast, RCTs randomly assign participants to experimental and control groups, maximizing control over confounding variables and being widely regarded as the gold standard for evaluating intervention effectiveness (Hariton & Locascio, 2018; Schulz et al., 2010). RCTs can be categorized as posttest-only or pretest–posttest designs. The pretest–posttest design examines individual changes before and after intervention, thereby controlling for individual differences to some extent and improving the precision of intervention effect estimation (Dugard & Todman, 1995), making it more scientifically rigorous than posttest-only designs. However, conventional pretest–posttest designs typically involve only a single measurement at each time point, which is vulnerable to situational factors, 偶然性, and measurement error, limiting precise estimation of intervention effects (Collins, 2006).

To further enhance validity, contemporary intervention research has increasingly shifted toward RCT designs with repeated measures at both pretest and posttest (RCT with repeated measures, RCTRM). This design combines the strengths of pretest–posttest and repeated measures approaches by implementing identical (or similar) repeated measurement protocols before and after intervention to track individual states multiple times. This approach enables researchers to repeatedly assess behavioral or psychological states (e.g., anxiety) in naturalistic settings or to collect dozens or even hundreds of experimental trials in laboratory settings to obtain stable response patterns. Such designs can capture changes in intra-individual stability while enhancing control over measurement error and

偶然性因素 (Bolger & Laurenceau, 2013). Consequently, RCTRM has become widely adopted in modern intervention research (e.g., Curran et al., 2010; Ferjan Ramírez et al., 2020; Gumisirizah et al., 2024; Shoenfelt et al., 2024).

1.2 Data Analysis in Intervention Research

Currently, several statistical approaches are commonly used for RCTRM data. First, traditional methods such as t-tests, ANOVA, and linear regression compare group means on intervention variables (e.g., anxiety, depression) at posttest or across pretest–posttest periods to evaluate intervention effects (e.g., Ferjan Ramírez et al., 2020; Kirby et al., 2023; Lago et al., 2021; Price et al., 2017; Thompson et al., 2024). Second, linear mixed-effects models (LMEM) or mixed-effects ANOVA account for the nested data structure (Aguilar-Raab et al., 2023; Barcaccia et al., 2024; Hanssen et al., 2023; Hellberg, 2021; Kreuder et al., 2020; Rees et al., 2015). In RCTRM designs, multiple trial outcomes are nested within participants, and this non-independence violates assumptions of traditional analytic methods (Aarts et al., 2014), leading to underestimated standard errors and inflated Type I error rates (Barr et al., 2013). Mixed models with random effects effectively address non-independence and yield accurate estimates.

However, existing statistical methods exhibit three notable limitations. First, they ignore heterogeneity of within-group variance (i.e., intra-individual variability, IIV). Although researchers increasingly use LMEM to capture between-individual differences, the homogeneity of residual variance assumption in traditional LMEM is frequently violated in practice. Ruscio and Roche (2021) demonstrated that variance heterogeneity is pervasive in psychological research, and ignoring it can cause group estimates to regress toward the overall mean (Williams, Mulder, et al., 2021), compromising exploration of true psychological characteristics and inflating Type I error rates (Walters et al., 2018). Second, evidence for intervention effects remains unidimensional. Current research primarily focuses on mean-level changes in intervention variables (McNeish & Matta, 2018), using the intervention variable itself as the outcome to provide evidence for intervention effectiveness. However, RCTRM designs with multiple measurement occasions offer opportunities to examine changes in intra-individual fluctuation (stability, i.e., IIV), such as learning trajectories (Williams et al., 2019) and temporal variation (Blozis et al., 2020), enabling multidimensional evaluation of intervention effects. Third, existing methods neglect moderating effects of covariates on intervention outcomes. As heterogeneity of intervention effects has gained attention, researchers increasingly recognize that intervention effects may vary across individuals (Kazdin, 2004; Sandler, Schoenfelder, Wolchik, & MacKinnon, 2011). Differences may exist across populations in pre-intervention status, fluctuations during intervention, and post-intervention outcomes. Therefore, it is essential to include individual-level covariates to examine how individual characteristics moderate intervention effects and enable personalized interventions. Current analytic approaches typically control for covariates to exclude their influence rather than examining their interactive effects

with intervention.

1.3 Analyzing Intervention Data with Mixed-Effects Location-Scale Models

The mixed-effects location-scale model (MELSM) extends traditional mixed-effects models by relaxing the assumption of homogeneous residual variance (Hedeker et al., 2008). In RCTRM designs, we assume a within-subject variable serves as the primary outcome for evaluating intervention effects. For example, in a study examining drug effects on anxiety, Y represents changes in anxiety states. The MELSM can separately model pretest MELSM and pretest-posttest difference MELSM.

Model 1: MELSM Without Covariates. Let y_{ij} denote the observed outcome for individual i (level 2) at trial j (level 1), where $i = 1, \dots, N$ and $j = 1, \dots, n_i$. N represents the number of participants, and n_i represents the number of trials for individual i. This study extends the random-intercept MELSM by decomposing the variance of outcome variable y_{ij} into between-individual and within-individual components for pretest and posttest phases.

Within-individual model: Pretest phase: $Y_{pre.ij} = \beta_{pre.0i} + e_{pre.ij}$ (1) Posttest phase: $Y_{post.ij} = \beta_{post.0i} + e_{post.ij}$ (2)

where $Y_{pre.ij}$ represents the outcome for participant i at trial j during the pretest phase, $Y_{post.ij}$ represents the outcome during the posttest phase, $\beta_{pre.0i}$ and $\beta_{post.0i}$ represent between-individual components before and after intervention, and $e_{pre.ij}$ and $e_{post.ij}$ represent within-individual components before and after intervention, following a multivariate normal distribution: $\begin{bmatrix} e_{pre.ij} \\ e_{post.ij} \end{bmatrix} \sim N(0, \Sigma_e = \begin{bmatrix} \sigma_{pre.i} & \sigma_{pre.post.i} \\ \sigma_{pre.post.i} & \sigma_{post.i} \end{bmatrix})$. The residual variance is allowed to vary between individuals, enabling evaluation of intra-individual variability (IIV) in intervention variables. Because variance is positive, we follow previous research by modeling the log of within-individual residual variance to capture IIV (Asparouhov et al., 2018; Hoffman & Walters, 2022).

Between-individual models separately construct pretest models and pretest-posttest difference models that reflect changes from pretest to posttest. The grouping variable uses dummy coding $G_i(0, 1)$ (0 = control group, 1 = intervention group). Intervention effects are reflected by regression coefficients of the grouping variable predicting the mean (i.e., $\beta_{pre.0i}$ and (b)) and intra-individual variability (i.e., $\sigma^2_{pre.i}$ and $\Delta \log(\sigma^2_i)$) in both pretest and pretest-posttest difference models.

Between-individual model: Pretest mean model: $\beta_{pre.0i} = \gamma_{00} + \gamma_{01} G_i + \mu_{0i}$ (3) Pretest variance model: $\sigma^2_{pre.i} = \exp(\omega_{00} + \omega_{01} G_i + \mu_{2i})$ (4) Pretest-posttest difference mean model: $(b) = \beta_{post.0i} - \beta_{pre.0i} = \gamma_{20} + \gamma_{21} G_i + u_{3i}$ (5) Pretest-

posttest difference variance model: $\Delta\log(\sigma_i^2) = \log(\sigma_{post.i}^2) - \log(\sigma_{pre.i}^2) = \omega_{10} + \omega_{11}G_i + u_{5i}$ (6)

where G_i indicates whether individual i belongs to the intervention or control group, γ_{00} and ω_{00} are fixed parameters in the pretest model representing the between-individual means of the outcome variable $\beta_{pre.0i}$ and residual variance $\sigma_{pre.i}^2$, respectively. γ_{01} and ω_{01} represent between-group differences in means and within-individual residuals in the pretest model, reflecting random assignment. γ_{20} and ω_{10} are fixed parameters in the pretest–posttest difference model representing between-individual means of changes in means and within-individual residuals from pretest to posttest. γ_{21} and ω_{11} represent effects of the grouping variable on mean differences and within-individual residual differences between pretest and posttest—i.e., intervention effects. Random parameters in the between-individual model represent unexplained between-individual differences, following a multivariate normal distribution.

Model 2: MELSM With Covariates. Building on Model 1, this version incorporates moderating variables at the individual level to examine moderating effects of individual covariates. Assuming an individual-level covariate W_i in the intervention, we examine whether intervention effects on means and IIV are moderated by this individual variable. For example, we might investigate whether social support moderates intervention effects on anxiety, such that individuals with higher support show better outcomes in both mean levels and IIV. Model 2' s within-individual model and pretest model are identical to Model 1 (equations 1-4). The between-individual model differs by adding the continuous individual covariate W_i to the pretest–posttest difference model, with moderating effects reflected by interaction terms between the covariate and grouping variable in predicting mean and IIV differences:

$(b) = \beta_{post.0i} - \beta_{pre.0i} = \gamma_{20} + \gamma_{21}G_i + \gamma_{22}W_i + \gamma_{23}G_{iW}i + u_{3i}$ (7) $\Delta\log(\sigma_i^2) = \log(\sigma_{post.i}^2) - \log(\sigma_{pre.i}^2) = \omega_{10} + \omega_{11}G_i + \omega_{12}W_i + \omega_{13}G_{iW}i + u_{5i}$ (8)

where W_i represents the value of the moderating variable for individual i, and parameters with the same notation as in Study 1 retain their previous meanings. γ_{22} and ω_{12} represent main effects of the moderating variable on mean and within-individual residual differences in the pretest–posttest difference model. γ_{23} and ω_{13} represent interaction effects between the grouping variable and moderating variable on mean and within-individual residual differences—i.e., moderating effects, which are the key parameters targeted in power analysis and effect size accuracy analysis.

This study first applies Model 1 (MELSM without covariates) to fit empirical data, demonstrating the necessity of using the pretest–posttest MELSM. Second, through Monte Carlo simulation studies of both Model 1 and Model 2, we investigate the performance of the pretest–posttest MELSM in detecting intervention and moderation effects and identify factors influencing this performance. Based on these findings, we provide recommendations and considerations for applying the pretest–posttest MELSM to RCTRM designs. All code used in this study

is available in the Appendix.

Empirical Example

The central aim of this study is to model intra-individual variability in intervention research by accounting for residual variance heterogeneity, thereby enabling multidimensional evaluation of intervention effects. To achieve this, we use the MELSM to develop a theoretical framework for intervention research (pretest–posttest MELSM). This section presents an empirical example to illustrate differences between evaluating interventions through mean levels versus intra-individual variability, and to demonstrate the importance of incorporating IIV as a novel outcome indicator.

2.1 Sample

The data used in this study come from a mindfulness intervention study, specifically the portion examining mindfulness effects on social media attractiveness. Participants were 65 college students aged 18 to 23 years (M = 20.05, SD = 1.46).

The study employed a randomized controlled trial design with repeated measures at both pretest and posttest, using a 2 (time: pretest, posttest) × 2 (group: intervention, control) mixed design. Participants were randomly assigned to a mindfulness painting group (N = 31) or a waitlist control group (N = 34). The mindfulness painting group received a 4-week mindfulness painting intervention, while the control group received only a one-hour lecture on mindfulness painting before the posttest. Both groups completed a repeated measurement task assessing social media attractiveness at pretest (using a two-choice oddball paradigm with 54 trials), which measured initial social media attractiveness (operationalized as reaction time differences compared to standard images). Within three days after the four mindfulness training sessions, both groups completed the same social media attractiveness measurement task at posttest. We fitted these data using both traditional repeated measures ANOVA and our Model 1 to demonstrate the advantages of the pretest–posttest MELSM.

2.3 Results

We first compared results from repeated measures ANOVA ([Figure 1: see original paper]) and pretest–posttest MELSM Model 1 (). ANOVA results showed no significant main effect of group, $F(1, 63) = 1.34$, $p = 0.252$, indicating no significant difference in overall reaction time between intervention and control groups. The main effect of time was marginally significant, $F(1, 63) = 3.46$, $p = 0.067$, suggesting a general decreasing trend in reaction time from pretest to posttest. The group × time interaction was marginally significant, $F(1, 63) = 3.90$, $p = 0.053$. Post-hoc comparisons revealed that within the intervention group, posttest reaction time was significantly lower than pretest, $t(63) = 2.71$, $p = 0.009$, whereas the control group showed no significant pretest–posttest

difference (p = 0.986). No significant group difference emerged at pretest (p = 0.999), but groups differed significantly at posttest, t(63) = 2.31, p = 0.024.

Traditional ANOVA thus suggested only a marginal intervention effect of mindfulness on social media attractiveness, providing no additional information about mindfulness intervention, such as its impact on within-individual fluctuations in social media attractiveness.

In contrast, Model 1 results () showed no significant between-group differences in social media attractiveness or its intra-individual variability at pretest ($\gamma_{01} = 0.659$, 95% CI = [-12.227, 13.258]; $\omega_{01} = 0.057$, 95% CI = [-0.097, 0.214]). No significant time effects emerged for either mean levels or IIV ($\gamma_{20} = -0.433$, 95% CI = [-9.355, 8.439]; $\omega_{10} = 0.040$, 95% CI = [-0.089, 0.167]). However, following mindfulness training, the intervention group showed significantly reduced social media attractiveness compared to the control group ($\gamma_{21} = -13.384$, 95% CI = [-25.927, -1.265]), indicating that mindfulness-trained individuals became less interested in social media. Additionally, the intervention group exhibited significantly reduced intra-individual variability in social media attractiveness ($\omega_{11} = -0.190$, 95% CI = [-0.372, -0.004]), suggesting that social media became a more stable (less variable) source of attraction after mindfulness training.

Using the pretest-posttest MELSM not only clarifies intervention effects on mean levels but also reveals changes in intra-individual variability (stability) of intervention variables. Furthermore, Model 2 can incorporate individual covariates to examine moderating effects, yielding richer and more detailed statistical results. Therefore, adopting MELSM is necessary for more accurate and nuanced data analysis, as traditional methods cannot provide these insights.

Based on Model 1 results, we created [Figure 2: see original paper], which displays random intercept variations for each participant before and after intervention. The x-axis represents participant IDs (sorted by pretest reaction time difference from low to high). Black dots indicate pretest point estimates, blue dots indicate posttest point estimates, and vertical lines represent credible intervals. Gray indicates cases where the credible interval includes the overall mean (no significant deviation from the population), while green and pink indicate cases where pretest and posttest credible intervals, respectively, exclude the overall mean (significant deviation). A higher proportion of colored vertical lines indicates greater individual differences from the population, warranting mixed-effects modeling.

The upper panels show mean model results (average reaction time differences), while lower panels show variance model results (individual IIV). In Panel A (control group), black (pretest) and blue (posttest) dashed lines nearly overlap, indicating no clear pretest-posttest difference in mean reaction time. Panel B (intervention group) shows posttest mean (blue dashed line) significantly lower than pretest (black dashed line), demonstrating a mean-level intervention effect. The positions of blue versus black dots reveal individual differences in intervention effects. Similarly, Panel C shows minimal pretest-posttest IIV differences

in the control group, indicating no time effect on IIV. Panel D (intervention group) shows posttest average IIV (blue dashed line) significantly lower than pretest (black dashed line), indicating reduced intra-individual variability after intervention. Comparing individual pretest-posttest IIV (blue vs. black dots) reveals that despite average IIV reduction, intervention effects on IIV still vary across individuals.

Moreover, comparing Panels B and D reveals that although individuals are sorted by mean reaction time difference, their intra-individual variability does not follow the same rank order. Larger means do not necessarily correspond to larger or smaller IIV. Only by including IIV as a dependent variable can reliable results be obtained. Therefore, using MELSM is essential for accounting for individual differences in both mean and variance models.

Study 1

Study 1 used Model 1 in an MCMC simulation study to examine the model' s performance in intervention research, focusing on statistical power and parameter estimation accuracy for γ_{21} and ω_{11} in the pretest-posttest difference model.

3.1 Data Generation Process

We generated data using Model 1. In the pretest mean model, the fixed intercept was set to $\gamma_{00} = 0$ (Walters et al., 2018). Assuming random assignment before intervention, the fixed slope for the grouping variable in the pretest mean model was $\gamma_{01} = 0$. In the pretest variance model, the fixed intercept was $\omega_{00} = 0$, and the residual variance for the mean structure was set to 1 (Walters et al., 2018). The fixed slope for the grouping variable in the pretest variance model was $\omega_{01} = 0$. In the pretest-posttest difference model, fixed intercepts were $\gamma_{20} = 0$ and $\omega_{10} = 0$. Based on comparisons of random effect estimates from the empirical example and pilot studies, random effect parameters τ_{33} and τ_{55} were fixed at 0.1. For simplicity, we constrained the two residuals to be independent (Walters et al., 2018) and fixed all correlations among random effects at 0 (Leckie et al., 2014).

3.2 Simulation Conditions

Study 1 employed a balanced design with equal numbers of participants in experimental and control groups. Simulation conditions included: (1) Level 1 sample size ($J = 20, 60, 100$); (2) Level 2 sample size ($I = 20, 60, 100$); (3) ICC effect size levels (0.1, 0.2, 0.3); (4) Pretest variance model random effect τ_{22} effect size levels (0.1, 0.3, 0.5); (5) Level 2 grouping variable G_i effect sizes (0, 0.2, 0.5, 0.8).

For conditions (1) and (2), we referenced previous simulation studies examining minimum sample sizes of 3 at Level 1 and 5 at Level 2 (Hecht & Zitzmann,

2021) and maximum sample sizes of 200 (Walters et al., 2018). Based on pilot study results, we selected $J = 20, 60, 100$ and $I = 20, 60, 100$. ICC refers to the intraclass correlation coefficient (Williams, Martin, et al., 2021), calculated as $ICC_i = \frac{\tau_{00}^2}{\tau_{00}^2+\sigma_i^2}$, where τ_{00}^2 represents the variance of the random intercept μ_{0i} in the pretest mean model and σ_i^2 represents residual variance. Since σ_i^2 varies in MELSM, ICC_i is also variable. For simplicity, we fixed σ_i^2 at its mean of 1 to calculate τ_{00}^2 (LeBreton & Senter, 2008). We set ICC effect size levels at 0.1, 0.2, and 0.3 (LeBreton & Senter, 2008), yielding random intercept effect sizes τ_{00}^2 of 0.111, 0.250, and 0.429.

For condition (4), we set pretest variance model random effect τ_{22} effect size levels at 0.1, 0.3, and 0.5, following previous research defining small, medium, and large random effects (Arend & Schäfer, 2019). For condition (5), we set Level 2 grouping variable G_i effect sizes at 0, 0.2, 0.5, and 0.8, referencing standard effect size levels for intervention effects (Zhou et al., 2021). An effect size of 0 indicates no intervention effect, used to assess Type I error rates. Because different variances with identical slopes can inflate power, we followed Arend et al. (2019) and used the formula $\gamma_{generate} = \gamma_{std}\sqrt{var}$, where var is the sum of all between-individual random effects, yielding $\gamma_{21} = \gamma_{std\_21}\sqrt{\tau_{33}^2 + \tau_{00}^2}$. Standardization for ω_{11} followed the same procedure. Final effect size indicators for parameters γ_{21} and ω_{11} across conditions are presented in .

These simulation factors were fully crossed, creating $3 \times 3 \times 4 \times 3 \times 3 = 324$ experimental condition combinations. For each combination, we generated 500 replications based on Model 1.

3.3 Analysis

Data generation and analysis were conducted using Mplus 8.10. Each dataset was fitted with the same model used to generate it (Model 1). We employed MCMC estimation based on Gibbs sampling (Chib, 1995) within a Bayesian framework to estimate posterior distributions, using posterior means as point estimates. Bayesian estimation used Mplus default non-informative priors: fixed regression coefficients $\gamma \sim N(0, +\infty)$ and random effect variances $\sigma^2 \sim \Gamma^{-1}(-1, 0)$. MCMC settings were: initial values = 0, number of chains = 2, iterations per chain = 10,000, warm-up = 5,000 iterations, thinning rate = 1. Convergence was assessed using the $\widehat{R} < 1.1$ criterion for all parameters (Williams, Liu, et al., 2021).

3.4 Evaluation Metrics

We comprehensively evaluated pretest-posttest MELSM performance using model convergence rates, statistical power, Type I error rates, and effect size estimation accuracy.

3.4.1 Convergence Rate Model convergence was examined across conditions, with acceptable convergence rates exceeding 90% (Li et al., 2024). Non-convergent replications were excluded from subsequent analyses:

$$convergence = \frac{N_{convergence}}{N}$$

where $N_{convergence}$ represents the number of successful convergences and N represents total simulation replications.

3.4.2 Statistical Power and Type I Error Rate Based on valid Monte Carlo results, we examined statistical power. Our focal parameters were γ_{21} and ω_{11}. An effect was considered correctly detected when the 95% credible interval excluded zero, indicating 95% probability that the parameter differed from zero. The desired power level was 0.8 (Cohen, 1988):

$$Power = \frac{N_{95\%CrI\ exclude\ zero}}{N_{convergence}}$$

When effect size was set to 0, this formula assessed Type I error rate—i.e., whether the model falsely detected intervention effects when none existed. Expected Type I error rates should range between 0.025 and 0.075.

3.4.3 Effect Size Estimation Accuracy and Standard Errors For our primary parameters γ_{21} and ω_{11}, we evaluated effect size estimation accuracy and standard error precision using:

$$Relative\ Bias = \frac{E(\hat{\theta})-\theta}{\theta} \quad (12) \quad RMSE = \sqrt{\frac{\sum(\hat{\theta}-\theta)^2}{N_{convergence}}} \quad (13) \quad 95\%CrI\ Coverage = \frac{N_{cover}}{N_{convergence}} \quad (14) \quad 95\%CrI\ Width = \frac{\theta_u-\theta_l}{N_{convergence}} \quad (15) \quad SE-SD\ bias = \widehat{SE}-SD_{\hat{\theta}} \quad (16)$$

where $E(\hat{\theta})$ is the expected value of parameter estimates (posterior means), θ is the true parameter value, N_{cover} is the number of replications where the 95% credible interval contained the true value, θ_u and θ_l are upper and lower credible interval bounds, $\widehat{SE}$ is the estimated standard error, and $SD_{\hat{\theta}}$ is the standard deviation of repeated parameter estimates.

Relative bias should be less than 0.1 in absolute value (Schultzberg & Muthén, 2018). RMSE should approach 0 (Schultzberg & Muthén, 2018). Coverage probability (95%CrI Coverage, CP) should range between 0.925 and 0.975 (Bradley, 1978). Credible interval width (95%CrI Width) reflects estimation precision; we used the difference between maximum and minimum effect sizes as the maximum acceptable width to encompass small, medium, and large effects (Maxwell et al., 2008). Width should be narrower than this maximum and 越小越好; and list maximum width standards across conditions. $SE-SD\ bias$ reflects standard error accuracy and should approximate 0 (Zhou et al., 2021).

3.5 Results

3.5.1 Convergence Rate The model converged across all simulation conditions, indicating good convergence properties suitable for further performance evaluation.

3.5.2 Type I Error Rate Type I error rates were below 0.075 across all conditions, indicating the model adequately controlled false positives ().

3.5.3 Statistical Power Power for the mean model parameter γ_{21} is shown in [Figure 3: see original paper]. Overall, power for γ_{21} increased with sample size, ICC, and intervention effect size. Random effect τ_{22} did not influence power for γ_{21}, consistent with expectations based on the ICC formula. Regarding sample size, Level 1 and Level 2 sample sizes were compensatory, but when Level 2 sample size was too small (e.g., $I = 20$), power never reached 0.8 regardless of Level 1 sample size. With small intervention effects (0.2), power standards could not be met unless ICC was large (0.3) and sample size was substantial ($I = 100$ and $J = 60$). With medium (0.5) to large (0.8) effects and medium (0.2) to large (0.3) ICC, power generally met standards except when sample size was minimal ($I = 20$ and $J = 20$).

Power for the variance model parameter ω_{11} is shown in [Figure 4: see original paper]. Overall, under equivalent conditions, power for ω_{11} was lower than for γ_{21}, indicating that detecting variance model intervention effects generally requires larger samples. For example, with small effect sizes (0.2), power never reached 0.8; supplementary analyses showed that sample sizes of 200 were needed to achieve adequate power for ω_{11}. Similar to the mean model, power for ω_{11} increased with sample size, intervention effect size, and random effect τ_{22}. ICC did not affect power for ω_{11}. Specifically, when Level 2 sample size was too small ($I = 20$), power never reached 0.8 regardless of Level 1 sample size. With small effects (0.2), no conditions met the 0.8 power criterion. With medium (0.5) to large (0.8) effects, adequate power required medium (0.3) to large (0.5) random effects τ_{22} and sufficient sample size ($I = 60$ and $J = 60$).

3.5.4 Parameter Estimation Accuracy presents relative bias results for both parameters under small intervention effect size (0.2). Overall, relative bias decreased as effect size increased; with medium (0.5) to large (0.8) effects, bias was acceptable, but small effects (0.2) produced overestimation. The variance model was more stringent than the mean model. Specifically, with small effects (0.2), γ_{21} met bias criteria except when ICC was small (0.1) and sample size was minimal ($I = 20$ and $J = 20$). For ω_{11}, overestimation occurred when τ_{22} was small (0.1) or medium (0.3) and Level 1 and Level 2 sample sizes were unbalanced (e.g., $J = 20$ and $I = 100$ or $J = 100$ and $I = 20$).

presents credible interval width results under medium intervention effect size (0.5). Maximum width standards varied by ICC and τ_{22} levels (see). Overall, effect size did not affect credible interval width; width narrowed as sample size

increased. The variance model was more demanding than the mean model. For γ_{21}, when ICC was small to medium (≤ 0.2), adequate sample size ($I = 60$ and $J = 60$) was required to meet width standards; when ICC was large (0.3), all conditions except minimal sample size ($I = 20$ and $J = 20$) met standards. For ω_{11}, when τ_{22} was small to medium (≤ 0.3), width standards were generally not met; only when τ_{22} was large (0.5) and sample size was substantial ($I = 100$ and $J = 100$) did results satisfy criteria.

RMSE, coverage, and standard error bias results appear in Supplementary Tables. Overall, RMSE and coverage for γ_{21} and ω_{11} aligned with relative bias patterns. Coverage probabilities fell between 0.925 and 0.975. Standard error bias ($SE - SD\ bias$) fluctuated near zero, indicating accurate standard error estimation. From a sample size planning perspective, to correctly identify and accurately estimate both intervention effects, researchers should use the maximum sample size required across both parameters. Under medium effect size (0.5), medium ICC (0.2), and medium random effect τ_{22} (0.3), a balanced design requires at least 100 participants and 100 trials.

Study 2

Building on Study 1, Study 2 incorporated an individual-level continuous covariate to examine its moderating effect on intervention effects. Using Monte Carlo simulation with Model 2, we investigated model performance in detecting moderation, focusing on power and parameter estimation accuracy for γ_{23} and ω_{13} in the pretest–posttest difference model.

4.1 Data Generation Process

We generated data using Model 2. Parameters shared with Model 1 retained their settings from Study 1. Because Study 2 focused on moderation rather than main intervention effects, we fixed intervention effects at medium levels: $\gamma_{21\_std} = 0.5$, $\omega_{11\_std} = 0.5$, with corresponding effect sizes calculated using the standardization formula (see). Similarly, we fixed covariate W_i main effects at medium levels (Arend & Schäfer, 2019): $\gamma_{22\_std} = 0.3$, $\omega_{12\_std} = 0.3$. The continuous covariate W_i was generated from a standard normal distribution $N(0, 1)$.

4.2 Simulation Conditions

Study 2 also used a balanced design with conditions: (1) Level 1 sample size ($J = 60, 100, 200$); (2) Level 2 sample size ($I = 60, 100, 200$); (3) ICC effect size levels (0.1, 0.2, 0.3); (4) Random effect levels (0.1, 0.3, 0.5); (5) Level 2 interaction effect sizes (0, 0.1, 0.3, 0.6).

For conditions (1) and (2), pilot studies indicated that moderation effects required larger sample sizes than main effects, so we expanded the range to $J = 60, 100, 200$ and $I = 60, 100, 200$. For condition (3), we examined how

ICC levels influenced covariate effects, setting small, medium, and large ICC levels at 0.1, 0.2, and 0.3 (LeBreton & Senter, 2008), yielding random intercept effect sizes τ_{00}^2 of 0.111, 0.250, and 0.429. For condition (4), we retained the same τ_{22} levels (0.1, 0.3, 0.5) as Study 1 to examine their influence on covariate effects. For condition (5), interaction effect sizes were set at 0, 0.1, 0.3, and 0.6 (Finch & French, 2011; Hoffman & Walters, 2022), representing small, medium, and large moderation effects. Corresponding effect sizes for γ_{23} and ω_{13} were calculated using the effect size formula with different ICC and random effect levels (see).

These factors were fully crossed, creating $3 \times 3 \times 3 \times 3 \times 4 = 324$ condition combinations. For each condition, we generated 500 replications based on Model 2.

4.3 Analysis

Because Mplus cannot generate data from models with covariates, Study 2 used R 4.2.3 for data generation and the MplusAutomation package to call Mplus for analysis and evaluation. Other settings matched Study 1. Evaluation metrics also remained consistent with Study 1.

4.4 Results

4.4.1 Convergence Rate All conditions achieved convergence, with 96.6% showing 100% convergence rates, suitable for further evaluation (see Supplementary Tables).

4.4.2 Type I Error Rate Type I error rates were acceptable (neither overly conservative nor inflated) across all conditions (see Appendix), indicating adequate control of false positives for moderation effects.

4.4.3 Statistical Power Power for the mean model moderation effect γ_{23} appears in [Figure 5: see original paper]. Overall, moderation effects required larger sample sizes than main effects to achieve equivalent power. Power for γ_{23} increased with sample size, moderation effect size, and ICC. Random effect τ_{22} did not influence power for γ_{23}. Level 1 and Level 2 sample sizes were compensatory, but power was sensitive to Level 2 sample size, requiring sufficient Level 2 participants. With small moderation effects (0.1) and small to medium ICC (0.1-0.2), power standards could not be met regardless of sample size; only with large ICC (0.3) and substantial sample size ($I = 200$ and $J = 200$) did power reach 0.8. With medium moderation effects (0.3), small to medium ICC (0.1-0.2) required large sample size ($I = 100$ and $J = 100$), whereas large ICC (0.3) met standards across all sample sizes. With large moderation effects (0.6), all conditions achieved adequate power.

Power for the variance model moderation effect ω_{13} appears in [Figure 6: see original paper]. Overall, variance model power was lower than mean model

power under equivalent conditions, and lower for moderation than for main effects. Power for ω_{13} increased with sample size, moderation effect size, and random effect τ_{22}. ICC did not influence power for ω_{13}. Specifically, small moderation effects (0.1) never met power standards. With medium effects (0.3), small to medium τ_{22} ($\leq$ 0.3) generally failed to meet standards; only large τ_{22} (0.5) with sufficient sample size ($I = 100$ and $J = 100$) achieved adequate power. With large effects (0.6), small τ_{22} (0.1) required sample size of 200, while medium to large τ_{22} (0.3–0.5) met standards except with minimal sample size ($I = 60$ and $J = 60$).

4.4.4 Parameter Estimation Accuracy presents credible interval width results under medium moderation effect size (0.5). Overall, moderation effect size did not affect width; width narrowed as sample size increased. The variance model was more stringent than the mean model. Specifically, with small ICC (0.1) and small Level 2 sample size ($I = 60$), width exceeded maximum standards. With medium (0.2) to large (0.3) ICC, all conditions met standards. With small random effect τ_{22} (0.1), no conditions met standards; with medium τ_{22} (0.3), sample size of 100 was required; with large τ_{22} (0.5), all conditions except minimal sample size ($I = 60$ and $J = 60$) satisfied criteria.

Relative bias, RMSE, coverage, and standard error bias results appear in Supplementary Tables. For moderation effects, γ_{23} relative bias met standards, indicating no systematic bias. For ω_{13}, relative bias was inflated due to extremely small true values (see), so we report raw bias, which was minimal. RMSE and coverage for γ_{23} and ω_{13} aligned with bias patterns. Coverage probabilities met standards. Standard error bias ($SE - SD\ bias$) fluctuated near zero. From a sample size planning perspective, to correctly identify and accurately estimate both moderation effects, researchers should use the maximum required sample size. Under medium moderation effect size (0.3), medium ICC (0.2), and medium random effect τ_{22} (0.3), a balanced design requires at least 200 participants and 100 trials.

Discussion and Conclusion

This study extends the mixed-effects location-scale model to RCTRM designs in intervention research. By incorporating intra-individual variability and enabling analysis of intervention effects under residual variance heterogeneity, we expand evaluation indicators in intervention research and demonstrate that IIV is an important outcome rather than mere "noise." This addresses previous limitations of unidimensional evaluation standards and biased results from ignoring within-group variance heterogeneity. The model can also include individual covariates to examine differential intervention effects across populations. Using power, effect size estimation, and accuracy metrics, two simulation studies examined factors influencing MELSM performance with and without covariates, confirming strong performance in RCTRM designs. Both intervention and moderation effects improved with larger effect sizes, sample sizes, ICC, and random

effect τ_{22} levels.

Findings indicate that sample size requirements for achieving adequate power in pretest–posttest MELSM exceed those for parameter estimation accuracy. Because model performance depends on both power and accuracy, we recommend considering both in study design, focusing on power, relative bias, and 95% credible interval width as key metrics. Under medium effect sizes with medium ICC and medium τ_{22}, the basic pretest–posttest MELSM requires at least 60 participants and 60 trials to detect mean intervention effects and 100 participants and 100 trials for IIV effects. For moderation effects, detecting mean moderation requires at least 60 participants and 100 trials, while IIV moderation requires 200 participants and 100 trials.

Regarding intervention effects, the variance model demands larger samples than the mean model. Results demonstrate that IIV can serve as a valuable intervention outcome, broadening evaluation indicators. Many studies have begun analyzing IIV changes, such as DSEM applications to autoregressive effects and IIV (e.g., McNeish & Hamaker, 2020). This study extends such analyses to broader repeated measures designs—specifically RCTRM. When examining IIV with small effect sizes (0.2), researchers must collect sufficient data; supplementary analyses showed that sample sizes of 200 were needed to detect variance model IIV effects. Therefore, studying IIV-related intervention effects requires increased trials or participants at the design stage. Although mean models require smaller samples, multidimensional evaluation of intervention effects justifies the additional sample size needed for variance models.

Regarding moderation effects, larger samples are required compared to main intervention effects. For example, under medium effect size (0.3), IIV intervention effects required sample sizes of 100, whereas IIV moderation effects required 200. The empirical example also showed that while 60 participants and 54 trials could detect intervention effects, they were insufficient for detecting moderation effects. This may be because anxiety (the moderator in the example) does not actually moderate mindfulness effects on social media attractiveness, or because the moderation effect was too small for the sample size to detect (Simulation Study 2 showed that small moderation effects remain undetectable even with $N = 200$). Indeed, moderation effects are typically smaller than main effects (Von Hippel & Schuetze, 2025). According to Von Hippel et al. (2025), interactions require caution and should only be examined when statistical power is sufficient. Therefore, increasing sample size yields more stable and reliable moderation effect estimates.

We offer the following recommendations for practitioners. First, if researchers seek to comprehensively evaluate intervention effects on both mean levels and intra-individual variability while accounting for nested data structures and variance heterogeneity, we recommend using our pretest–posttest MELSM. Second, for proper identification and accurate estimation of intervention effects, we advise conducting simulation-based sample size planning before research. This involves power and accuracy analyses with parameter values including fixed

and random effects for both outcome dimensions (means and IIV). We suggest referencing previous MELSM simulation studies for parameter settings (e.g., Walters et al., 2018) and using conventional effect size benchmarks (e.g., Cohen' s d) or meta-analytic results from relevant intervention research for fixed effects. When prior research is unavailable, pilot studies can inform parameter settings, though biased pilot estimates may yield unreasonable sample sizes (Albers & Lakens, 2018). Alternative approaches include sampling true values from pilot effect size distributions to account for uncertainty and obtain more reasonable sample size ranges. Researchers may also directly reference our sample size recommendations. Regardless, simulation studies should explicitly justify all parameter settings to standardize sample size planning procedures.

Limitations and Future Directions

Several limitations remain. First, both simulation studies used balanced designs with equal group sizes, whereas real-world research often involves unbalanced designs. Future research should extend to unbalanced designs to examine intervention effects in special contexts where adequate control group matches cannot be obtained, necessitating performance comparisons and sample size planning for unbalanced scenarios. Second, this study only considered individual-level covariates (Level 2) without within-individual covariates (Level 1 slope models). Future research could incorporate within-individual covariates to explore additional psychological questions, such as examining trial-level effects in experimental designs where trials are nested within individuals.

References

Aarts, E., Verhage, M., Veenvliet, J. V., Dolan, C. V., & Van Der Sluis, S. (2014). A solution to dependency: Using multilevel analysis to accommodate nested data. *Nature Neuroscience, 17*(4), 491–496. https://doi.org/10.1038/nn.3648

Aguilar-Raab, C., Winter, F., Warth, M., Stoffel, M., Moessner, M., Hernández, C., Pace, T. W. W., Harrison, T., Negi, L. T., Jarczok, M. N., & Ditzen, B. (2023). A compassion-based treatment for couples with the female partner suffering from current depressive disorder: A randomized-controlled trial. *Journal of Affective Disorders*. https://doi.org/10.1016/j.jad.2023.08.136

Albers, C., & Lakens, D. (2018). When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *Journal of Experimental Social Psychology, 74*, 187–195. https://doi.org/10.1016/j.jesp.2017.09.004

Arend, M. G., & Schäfer, T. (2019). Statistical power in two-level models: A tutorial based on Monte Carlo simulation. *Psychological Methods, 24*(1), 1–19. https://doi.org/10.1037/met0000195

Asparouhov, T., Hamaker, E. L., & Muthén, B. (2018). Dynamic Structural Equation Models. *Structural Equation Modeling: A Multidisciplinary Journal,*

25(3), 359–388. https://doi.org/10.1080/10705511.2017.1406803

Barcaccia, B., Medvedev, O. N., Pallini, S., Mastandrea, S., & Fagioli, S. (2024). Examining Mental Health Benefits of a Brief Online Mindfulness Intervention: A Randomised Controlled Trial. *Mindfulness, 15*(4), 835–843. https://doi.org/10.1007/s12671-024-02331-8

Blozis, S. A., McTernan, M., Harring, J. R., & Zheng, Q. (2020). Two-part mixed-effects location scale models. *Behavior Research Methods, 52*(5), 1836–1847. https://doi.org/10.3758/s13428-020-01384-4

Bolger, N., & Laurenceau, J.-P. (2013). *Intensive Longitudinal Methods: An Introduction to Diary and Experience Sampling Research*. Guilford Press.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*(2), 144–152. https://doi.org/10.1111/j.2044-8317.1978.tb00581.x

Brailovskaia, J., Teismann, T., & Margraf, J. (2020). Physical activity mediates the association between daily stress and Facebook Addiction Disorder (FAD)–A longitudinal approach among German students. *Computers in Human Behavior, 110*, 106355. https://doi.org/10.1016/j.chb.2019.106355

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association, 90*(432), 1313–1321. https://doi.org/10.1080/01621459.1995.10476635

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum Associates.

Collins, L. M. (2006). Analysis of Longitudinal Data: The Integration of Theoretical Model, Temporal Design, and Statistical Model. *Annual Review of Psychology, 57*(1), 505–528. https://doi.org/10.1146/annurev.psych.57.102904.190146

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Houghton Mifflin.

Curran, P. J., Obeidat, K., & Losardo, D. (2010). Twelve Frequently Asked Questions About Growth Curve Modeling. *Journal of Cognition and Development, 11*(2), 121–136. https://doi.org/10.1080/15248371003699969

Dugard, P., & Todman, J. (1995). Analysis of Pre-test-Post-test Control Group Designs in Educational Research. *Educational Psychology, 15*(2), 181–198. https://doi.org/10.1080/0144341950150207

Ferjan Ramírez, N., Lytle, S. R., & Kuhl, P. K. (2020). Parent coaching increases conversational turns and advances infant language development. *Proceedings of the National Academy of Sciences, 117*(7), 3484–3491. https://doi.org/10.1073/pnas.1921653117

Finch, W. H., & French, B. F. (2011). Estimation of MIMIC Model Parameters with Multilevel Data. *Structural Equation Modeling: A Multidisciplinary Journal, 18*(2), 229–252. https://doi.org/10.1080/10705511.2011.557338

Gumisirizah, N., Nzabahimana, J., & Muwonge, C. M. (2024). Students' performance, attitude, and classroom observation data to assess the effect of problem-based learning approach supplemented by YouTube videos in Ugandan classroom. *Scientific Data, 11*(1), 428. https://doi.org/10.1038/s41597-024-03206-2

Hanssen, I., Huijbers, M., Regeer, E., Lochmann Van Bennekom, M., Stevens, A., Van Dijk, P., Boere, E., Havermans, R., Hoenders, R., Kupka, R., & Speckens, A. E. (2023). Mindfulness-based cognitive therapy v. treatment as usual in people with bipolar disorder: A multicentre, randomised controlled trial. *Psychological Medicine, 53*(14), 1–12. https://doi.org/10.1017/S0033291723000090

Hariton, E., & Locascio, J. J. (2018). Randomised controlled trials—The gold standard for effectiveness research. *BJOG: An International Journal of Obstetrics and Gynaecology, 125*(13), 1716. https://doi.org/10.1111/1471-0528.15199

Hecht, M., & Zitzmann, S. (2021). An Introduction to Multilevel Modeling for Experimental Psychology. *Experimental Psychology, 68*(3), 131–143. https://doi.org/10.1027/1618-3169/a000522

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. https://doi.org/10.1111/j.1541-0420.2007.00924.x

Hellberg, S. N. (2021). Extending cognitive-behavioral theory with ecological momentary assessment: An application to panic disorder. *Behavior Research and Therapy, 143*, 103847. https://doi.org/10.1016/j.brat.2021.103847

Hoffman, L., & Walters, R. W. (2022). Catching Up on Multilevel Modeling. *Annual Review of Psychology, 73*(1), 659–689. https://doi.org/10.1146/annurev-psych-020821-103525

Kazdin, A. E. (2004). Evidence-based treatments: Challenges and priorities for practice and research. *Child and Adolescent Psychiatric Clinics of North America, 13*(4), 923–940. https://doi.org/10.1016/j.chc.2004.05.001

Kazdin, A. E. (2017). *Research Design in Clinical Psychology* (5th ed.). Pearson.

Kirby, J. N., Hoang, A., & Ramos, N. (2023). A brief compassion focused therapy intervention can help self-critical parents and their children: A randomised controlled trial. *Psychology and Psychotherapy: Theory, Research and Practice, 96*(3), 1–18. https://doi.org/10.1111/papt.12459

Kreuder, A.-K., Scheele, D., Schultz, J., Hennig, J., Marsh, N., Dellert, T., Ettinger, U., Philipsen, A., Babasiz, M., Herscheid, A., Remmersmann, L., Stirnberg, R., Stöcker, T., & Hurlemann, R. (2020). Common and dissociable effects of oxytocin and lorazepam on the neurocircuitry of fear. *Proceedings of the National Academy of Sciences, 117*(21), 11781–11787. https://doi.org/10.1073/pnas.1920147117

Lago, T. R., Brownstein, M. J., Page, E., Beydler, E., Manbeck, A., Beale, A., Roberts, C., Balderston, N., Damiano, E., Pineles, S. L., Simon, N., Ernst, M., & Grillon, C. (2021). The novel vasopressin receptor (V1aR) antagonist SRX246 reduces anxiety in an experimental model in humans: A randomized proof-of-concept study. *Psychopharmacology, 238*(9), 2393–2403. https://doi.org/10.1007/s00213-021-05861-4

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 Questions About Interrater Reliability and Interrater Agreement. *Organizational Research Methods, 11*(4), 815–852. https://doi.org/10.1177/1094428106296642

Leckie, G., French, R., Charlton, C., & Browne, W. (2014). Modeling Heterogeneous Variance–Covariance Components in Two-Level Models. *Journal of Educational and Behavioral Statistics, 39*(5), 307–332. https://doi.org/10.3102/1076998614546494

Li, Y., Williams, L., Muth, C., Heshmati, S., Chow, S.-M., & Oravecz, Z. (2024). A Growth of Hierarchical Autoregression Model for Capturing Individual Differences in Changes of Dynamic Characteristics of Psychological Processes. *Structural Equation Modeling: A Multidisciplinary Journal, 31*(4), 1–14. https://doi.org/10.1080/10705511.2024.2402328

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample Size Planning for Statistical Power and Accuracy in Parameter Estimation. *Annual Review of Psychology, 59*(1), 537–563. https://doi.org/10.1146/annurev.psych.59.103006.093735

McNeish, D., & Hamaker, E. L. (2020). A primer on two-level dynamic structural equation models for intensive longitudinal data in Mplus. *Psychological Methods, 25*(5), 610–635. https://doi.org/10.1037/met0000250

McNeish, D., & Matta, T. (2018). Differentiating between mixed-effects and latent curve approaches to growth modeling. *Behavior Research Methods, 50*(4), 1398–1414. https://doi.org/10.3758/s13428-017-0976-3

Price, D., Burris, D., Cloutier, A., Thompson, C. B., Rilling, J. K., & Thompson, R. R. (2017). Dose-Dependent and Lasting Influences of Intranasal Vasopressin on Face Processing in Men. *Frontiers in Endocrinology, 8*, 220. https://doi.org/10.3389/fendo.2017.00220

Rees, C. S., Hasking, P., Breen, L. J., Lipp, O. V., & Mamotte, C. (2015). Group mindfulness based cognitive therapy vs group support for self-injury among young people: Study protocol for a randomised controlled trial. *BMC Psychiatry, 15*(1), 154. https://doi.org/10.1186/s12888-015-0525-9

Ruscio, J., & Roche, B. (2021). Variance heterogeneity in psychological research: A review and demonstration of its importance. *Psychological Methods, 27*(3), 401–418. https://doi.org/10.1037/met0000370

Sandler, I. N., Schoenfelder, E. N., Wolchik, S. A., & MacKinnon, D. P. (2011). Long-term impact of prevention programs to promote effective parenting: Last-

ing effects but uncertain processes. *Annual Review of Psychology, 62*, 299–329. https://doi.org/10.1146/annurev.psych.121208.131619

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. https://doi.org/10.1080/10705511.2017.1392862

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin.

Shoenfelt, A., Pehlivanoglu, D., Lin, T., Ziaei, M., Feifel, D., & Ebner, N. C. (2024). Effects of chronic intranasal oxytocin on visual attention to faces vs. Natural scenes in older adults. *Psychoneuroendocrinology, 164*, 107018. https://doi.org/10.1016/j.psyneuen.2024.107018

Thompson, R. R., Price, D., Burris, D., Cloutier, A., & Rilling, J. K. (2024). Effects of arginine vasopressin on human anxiety and associations with sex, dose, and V1a-receptor genotype. *Psychopharmacology, 241*(6), 1177–1190. https://doi.org/10.1007/s00213-024-06551-7

Von Hippel, P. T., & Schuetze, B. A. (2025). How Not to Fool Ourselves About Heterogeneity of Treatment Effects. *Advances in Methods and Practices in Psychological Science, 8*(2), 25152459241304347. https://doi.org/10.1177/25152459241304347

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), 1–23. https://doi.org/10.1080/00273171.2018.1449628

Williams, D. R., Liu, S., Martin, S. R., & Rast, P. (2021). Bayesian Multivariate Mixed-Effects Location Scale Modeling of Longitudinal Relations among Affective Traits, States, and Physical Activity. *Psychological Assessment, 33*(8), 1–14. https://doi.org/10.1037/pas0001037

Williams, D. R., Martin, S. R., & Rast, P. (2021). Putting the individual into reliability: Bayesian testing of homogeneous within-person variance in hierarchical models. *Behavior Research Methods, 54*(3), 1272–1290. https://doi.org/10.3758/s13428-021-01646-x

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. https://doi.org/10.1037/met0000270

Williams, D. R., Zimprich, D. R., & Rast, P. (2019). A Bayesian nonlinear mixed-effects location scale model for learning. *Behavior Research Methods, 51*(5), 1–18. https://doi.org/10.3758/s13428-019-01255-9

Zhou, L., Wang, M., & Zhang, Z. (2021). Intensive Longitudinal Data Analyses With Dynamic Structural Equation Modeling. *Organizational Research Methods, 24*(2), 219–250. https://doi.org/10.1177/1094428119833164

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