METHODOLOGY
European Journal of Research Methods for the Behavioral and Social Sciences

PsychOpen GOLD

Original Article

Check for updates

Bayesian Versus Frequentist Approaches in Multilevel Single-Case Designs: On Type I Error Rate and Power

Cristina Rodríguez-Prada[1], José Ángel Martínez-Huertas[2], Ricardo Olmos[1]

[1] *Department of Social Psychology and Methodology, School of Psychology, Universidad Autónoma de Madrid, Madrid, Spain.* **[2]** *Department of Methodology of Behavioral Sciences, School of Psychology, Universidad Nacional de Educación a Distancia, Madrid, Spain.*

Methodology, 2026, Vol. 22(1), 52–76, https://doi.org/10.5964/meth.17715

Received: 2025-04-16 • **Accepted:** 2026-01-13 • **Published (VoR):** 2026-03-27

Handling Editor: Tamás Rudas, Eötvös Loránd University, Budapest, Hungary

Corresponding Author: Cristina Rodríguez-Prada, "Aula PDIF", Department of Social Psychology and Methodology, Universidad Autónoma de Madrid (UAM), Madrid, Spain.

Supplementary Materials: Code, Data, Materials [see Index of Supplementary Materials]

Abstract

Single-case designs (SCEDs) assess intervention effects through repeated measurements on one or a few individuals. Multilevel models nest repeated measures within individuals and have gained popularity for inferential analysis in SCEDs, in combination with expert knowledge of the clinicians and applied researchers. However, researchers often face model specification challenges without knowing the true population model underlying their data. This study evaluates how model selection criteria (AIC, BIC, WAIC, LOO) conditioned on the selected model impact statistical power and Type I error rates in intervention effects, reflecting the ecological reality where practitioners do not know the true model. A Monte Carlo simulation modelled data of AB designs varying sample size, measurement points, intervention effects, and random effect structures. Competing multilevel models were then fitted and compared using AIC, BIC, WAIC, and LOO to examine the impact of model selection on statistical power and Type I error rates. Results indicated that frequentist criteria performed well in simpler models in terms of power, while Bayesian approaches showed greater robustness with respect to Type I error control. The findings provide practical insights on multilevel model selection under real-world conditions, highlighting Bayesian methods as a robust alternative for applied researchers handling small sample sizes and complex data structures.

Keywords

single-case designs, multilevel analysis, Bayesian statistics, frequentist analysis, statistical power, Type I Error rate

Single-case experimental designs (SCEDs) provide a valuable framework for analysing intervention effects on individuals through repeated measures (Bono & Arnau, 2014; Kazdin, 1982; Shadish & Sullivan, 2011). Their objective is to establish a functional relationship between an intervention and changes in a behavioural outcome, as applied contexts often hinder meeting traditional causal criteria (logical connection, covariation, temporal precedence, and full control of confounding variables; Kazdin, 1977; Manolov et al., 2014; Virués-Ortega & Haynes, 2005). While traditional qualitative methods, like visual inspection of time-series data, can be somewhat subjective and less effective for detecting subtle changes (Busk & Serlin, 1992; Kazdin, 1982; Kratochwill et al., 2013; Manolov & Moeyaert, 2017; Parsonson & Baer, 1986; Van den Noortgate & Onghena, 2003a), quantitative methods, particularly multilevel models, have increasingly addressed these considerations.

Multilevel linear models (MLMs) address the nested structure of the data, improving the precision of effect estimates and enabling the analysis of contextual variables at multiple levels (Hoffman, 2014; Manolov & Moeyaert, 2017; Moeyaert et al., 2020; Van den Noortgate & Onghena, 2003b). They effectively address statistical issues like autocorrelation, which are inherent to SCED data and can undermine the validity of parametric tests (Bono & Arnau, 2014; Gentile et al., 1972; Keselman & Leventhal, 1974). MLMs also complement non-parametric methods in SCED analysis, such as non-overlap indices (PND, PEM, NAP; Parker et al., 2011), which measure improvement across phases but have certain limitations (Rodríguez-Prada & Olmos, 2019). MLMs provide a comprehensive framework for statistical inference in SCEDs (Botella & Caperos, 2019).

MLMs can overcome some SCED challenges, such as model specification issues, limited information for estimating parameters and incorrect standard error estimations from techniques that overlook the data's hierarchical structure (Rodabaugh & Moeyaert, 2017). By modelling hierarchical structures, MLMs efficiently capture trends and allow flexible error covariance specifications, enhancing robustness and reliability in statistical decisions (Hoffman, 2014). MLMs also analyse fixed effects, shared among participants, and random effects, which account for individual variability in treatment outcomes (Manolov & Moeyaert, 2017). This dual-level approach is particularly well-suited for SCEDs, where individual responses are central to understanding intervention effects, and the hierarchical nature of the data must be accounted for to ensure valid conclusions (Moeyaert et al., 2014; Van den Noortgate & Onghena, 2003a; 2003b). Researchers can assess individual factors influencing clinical outcomes by measuring between-individual variability by explicitly modelling moderators to address why some benefit more from interventions (Moeyaert et al., 2024; Moeyaert & Yang, 2021). MLMs handle complex data structures, specify covariance structures, and model variability, making them suitable

for small-sample designs including SCEDs. However, challenges remain in estimating higher-level variance components with few individuals.

Two main approaches to estimate MLMs are frequentist and Bayesian frameworks, differing in parameter estimation and uncertainty quantification. Frequentist methods, typically using maximum likelihood estimation (ML), view parameters as fixed but unknown. Restricted Maximum Likelihood (REML) is preferred for random effects models, as it adjusts for the loss of degrees of freedom in variance component estimation. However, these approaches depend on asymptotic assumptions that require large samples for accuracy (Hoffman, 2014). In small-sample contexts like SCEDs, these assumptions often fail, resulting in biased covariance estimates despite unbiased fixed effect estimates (Baek et al., 2020; Moeyaert et al., 2017; Van den Noortgate & Onghena, 2003a).

Bayesian estimation has emerged as a promising alternative for small-sample designs in SCED studies (Baek et al., 2020; McNeish, 2016). Unlike frequentist methods, it incorporates prior information about parameters, helping to mitigate limitations of datasets with few Level-2 units (Gelman et al., 2013; van de Schoot et al., 2015). Selecting appropriate prior distributions is crucial in estimating Level-2 variance components, as weak or overly restrictive priors can impact results (Baek & Ferron, 2020; Moeyaert et al., 2017). Balancing informative and weakly informative priors is essential to prevent bias and ensure robustness. Additionally, Bayesian methods use advanced algorithms like Markov Chain Monte Carlo (MCMC; Brooks, 1998) to estimate parameters in complex models, enhancing precision and stability (McNeish, 2016; Rindskopf, 2014).

Whether using frequentist or Bayesian frameworks, a major challenge in applying MLMs to SCED data is selecting the best model and its random effect structure. This choice affects the balance between Type I error rates and statistical power, influencing hypothesis testing conclusions (Matuschek et al., 2017). Underparameterized models may oversimplify data and miss key patterns, while overparameterized models can inflate standard errors and reduce statistical power (Hoffman, 2014; Martínez-Huertas et al., 2022; Martínez-Huertas & Olmos, 2022). Thus, model selection is essential for valid and meaningful statistical decisions in SCED data interpretation, in combination with expert knowledge of the clinicians and applied researchers.

A practical approach uses information criteria to compare models, rewarding good fits and penalising overfitting. These criteria arise from Kullback-Leibler (KL) divergence and entropy, estimating model generalizability by prioritising predictive accuracy over simple goodness of fit. Frequentist methods employ the Akaike Information Criterion (AIC; Akaike, 1998) and Bayesian Information Criterion (BIC; Schwarz, 1978), with AIC favouring models that minimise information loss and BIC selecting the most probable model under a Bayesian framework with stricter complexity penalties (Raftery, 1995; Weakliem, 1999). Bayesian approaches use indices like the Watanabe-Akaike Information Criterion (WAIC; Watanabe, 2010) and Leave-One-Out Cross-Validation (LOO; Vehtari et al., 2017) to enhance model evaluation by incorporating posterior distributions. In con-

trast to frequentist criteria, WAIC prioritises predictive utility over determining the true population model (Nicenboim & Vasishth, 2016), while LOO assesses model performance using cross-validation across data points, which makes it robust to outliers. Both indices select the model with the lowest value, and WAIC and LOO often outperform AIC and Deviance Information Criterion (DIC; Spiegelhalter et al., 2002), offering greater reliability in comparisons (Gelman et al., 2013). In the present simulation study, the performance of all indices is assessed according to their ability to identify the data-generating model.

Aims of the Present Study

This study compares Bayesian and frequentist methods in MLMs for model selection in SCED data via simulation. It focuses on the information criteria AIC, BIC, WAIC, and LOO fit indices to select the true population model and analyze their impact on statistical power and Type I errors. Three objectives are established. The first objective is to evaluate if various Bayesian priors present differences regarding statistical power and Type I error rates. Building on Moeyaert et al. (2017) and Baek et al. (2020), we explore how prior specifications affect robustness and accuracy. The second objective assesses selection accuracy using frequentist (AIC, BIC) and Bayesian (WAIC, LOO) information criteria, aiming for consistency in the Bayesian framework. The third examines how model selection affects statistical power and Type I error rates in intervention effects, identifying optimal estimation methods under unknown population models. In this context, we evaluate the operating characteristics (Type I error and power) of different model selection procedures (AIC, BIC, WAIC, LOO) when the data-generating process is unknown. Thus, this study addresses the gap of analysing model selection's impact from an ecological viewpoint, reflecting researchers' experiences in applied settings. In practical terms, this paper examines how AIC, BIC, WAIC, and LOO behave in terms of statistical power and Type I error rates when the true data-generating multilevel model is unknown — a common situation in applied SCED studies. Based on these results, the study aims to provide an evidence-based recommendation on which criterion offers the best trade-off between power and Type I error control under model uncertainty.

Method

Simulation Study Design

A Monte Carlo simulation of an SCED AB-design with baseline (A) and intervention (B) phases was conducted. The dependent variables included statistical power, Type I error rate, and proportion of true model selection per information criteria. Statistical power indicates the rate of correctly detecting the intervention effect, while the Type I error rate measures the incorrect identification of an effect when none exists. Correct model

selection quantifies how often the information criteria identified the true population model during comparison.

Data Generation and Population Values

SCEDs data were generated based on four models varying in the number of random effects (covariance parameters). The models were:

1. **Minimal model**: Assumes a constant for the initial value and the intervention effect across individuals. This means that the dependent variable is initially absent or at a floor level for all individuals, and that the intervention effect does not present variation between individuals:

$$Y_{ij} = \gamma_{00} + \gamma_{10} * condition_{ij} + e_{ij},$$

where Y_{ij} is the response variable for individual j at observation i, γ_{00} represents baseline mean (intercept), γ_{10} is the average intervention effect (slope), e_{ij} is the Level-1 residual error for individual j at observation i and $condition_{ij}$ is a dummy-coded predictor indicating whether observation i for individual j belongs to the baseline phase (0) or the intervention phase (1). This model does not include random effects.

2. **Partial intercepts model:** Includes random intercepts for baseline differences among individuals. This model is relevant in SCED contexts where individuals have different baseline levels, but the intervention is expected to produce a uniform effect:

$$Y_{ij} = \gamma_{00} + \gamma_{10} \cdot condition_{ij} + u_{0j} + e_{ij},$$

where u_{0j} is the random intercept in the baseline for individual j.

3. **Partial slopes model:** Incorporates random slopes to represent different intervention effects among individuals, essential in SCED contexts where responses are initially low or absent, or where individual characteristics may influence intervention outcomes:

$$Y_{ij} = \gamma_{00} + \gamma_{10} \cdot condition_{ij} + u_{1j} \cdot condition_{ij} + e_{ij},$$

where u_{1j} is the random slope for individual j.

4. **Maximal model**: Includes both random intercepts and slopes, accounting for variability in both baseline scores and intervention effects. This is the most ecological model because it considers the variability that an individual's learning history and personal factors can hold:

$$Y_{ij} = \gamma_{00} + \gamma_{10} \cdot condition_{ij} + u_{0j} + u_{1j} \cdot condition_{ij} + e_{ij},$$

where u_{0j} and u_{1j} are random intercepts and slopes for individual j, respectively.

In all the population models, fixed effects were set at $\gamma_{10} = 5$, random effects followed normal distributions $u_0 \sim N(0, 2)$, $u_1 \sim N(0, 2)$, and residual variance was set to $N(0, 1)$. Random effects were uncorrelated ($r = 0$) to minimise model complexity, as suggested by Van den Noortgate and Onghena (2003a) and Moeyaert et al. (2017). In addition, residuals were assumed to be independent over time, and no autocorrelation or secular nor temporal trends were simulated.

Simulation Conditions

The simulation comprised 144 scenarios from various factor combinations:

1. Number of individuals (N_j = 3, 5, 7) (Moeyaert et al., 2017; Shadish & Sullivan, 2011).

2. Number of repeated measurements ($RRMM_j$ = 10, 20, 30, 40) (Moeyaert et al., 2017, Shadish & Sullivan, 2011). We added two additional values (10 and 30) to increase design resolution across short-to-moderate series lengths and to assess whether performance changes monotonically with series length.

3. Effect size for intervention effect γ_{10}: null ($d = 0$), medium ($d = 1.15$), and large ($d = 2.70$), following Ferguson's (2009) guidelines.

4. Population model structure, ranging from a Minimal model (no random effects) to increasingly complex specifications (Partial Intercepts, Partial Slopes, and Maximal models). These random structures are in line with previous simulation studies (Martínez-Huertas et al., 2022; Martínez-Huertas & Olmos, 2022; Matuschek et al., 2017).

In each scenario, four model information criteria were evaluated: frequentist criteria (AIC, BIC) and Bayesian criteria (WAIC, LOO), allowing a comparative analysis of selection approaches. Seven different priors were tested to assess their effects on statistical power, Type I error rates for the intervention effect, and the accuracy of model selection using WAIC and LOO. In line with the default parameterisation in *brms*, weakly informative priors (see Gelman, 2006; McElreath, 2020) were specified on the standard deviations of the random effects (σ_{u0} and σ_{u1}), not on the variance components (σ^2_{u0} and σ^2_{u1}). Because standard deviations are constrained to be positive, Half-Cauchy and Half-normal priors with scale parameters of 10, 20, and 50 were used, alongside a weakly informative Uniform (0, 100) prior. Following the approach outlined by Moeyaert et al. (2017), who derived these values from reanalyses of empirical SCED studies with normally distributed continuous outcomes, our aim was to reflect plausible ranges for variance components in this context. This approach allows the priors to regularize estimation in small-sample settings while still letting the normally distributed outcome data contribute substantially to the results. Increasing the scale parameter in the Half-Cauchy and Half-Normal distributions makes the prior less informative by allowing for greater variability in the variance components. A smaller scale concentrates the prior mass closer to zero, while a larger scale spreads the distribution and assigns more weight to larger variance

values, thus reducing the prior's influence. We adopted this approach because one of the main challenges in this context is selecting the appropriate prior distribution. Although future studies should explore this issue in greater depth, our current aim is to increase variability so that the prior parameterization remains sensible while still allowing the data to provide valuable insight.

Conducting the Simulation Study

Five hundred replications were conducted for each of the 144 simulation conditions in R. Frequentist models used the *lme4* package (Bates et al., 2015) for models with random effects and *nlme* package (Pinheiro et al., 2021) for the minimal model, always using REML as the estimation method and Satterthwaite's approximation to the denominator degrees of freedom (via *lmerTest*). Bayesian estimation utilised the *brms* package (Bürkner, 2017), employing the Hamiltonian Monte Carlo (HMC) algorithm with its NUTS extension for efficient sampling of complex models (Hoffman & Gelman, 2014). Bayesian models were fitted using two chains of 1,000 iterations each, with 400 warm-up iterations per chain, yielding a total of 1,200 post–warm-up draws per model. The *adapt_delta* parameter was set to 0.95 to reduce divergent transitions and enhance sampling reliability. Extracted data included point estimates, standard deviations, p values (frequentist), and posterior means with 95% credible intervals (Bayesian). Information Criteria fit indices (AIC, BIC, WAIC, LOO) and convergence diagnostics were also extracted ($\widehat{R}$, with $\widehat{R} < 1.1$ indicating satisfactory convergence). According to this criterion, approximately 99% of the replicas showed appropriate convergence. The convergence dataset is available on our OSF project (see Rodríguez-Prada et al., 2026a).

Data Analysis

Data processing and analysis were performed in R using RStudio. Power and Type I error rates were calculated based on intervention effects. Frequentist decisions used a significance threshold ($p < 0.05$), while Bayesian ones relied on 95% credible intervals (CIs). Model selection was assessed using AIC, BIC, WAIC, and LOO by comparing each candidate model (minimal, partial intercepts, partial slopes, maximal) to the population model. In addition, analysis of variance (ANOVA) was conducted to examine the influence of simulation parameters on performance indices, and partial eta-squared (η_p^2) was reported as a measure of effect size. Partial eta-squared was used as a relative indicator of the impact of simulation conditions. Effect size magnitudes were not interpreted using conventional cut-off values, as the aim was not to compare absolute effect sizes with previous studies but to examine differences across simulation conditions. Partial eta-squared was selected over eta-squared because it provides an unbiased estimate of the unique variance explained by each factor in multifactorial designs (Richardson, 2011). Consistent with prior methodological simulation studies (e.g., Moeyaert et al., 2017), cut-off points

of .01, .06, and .14 were adopted to classify small, medium, and large effects, respectively (Cohen, 1988). This approach allowed us to identify the most influential simulation parameters beyond statistical significance. The dataset, prepared for the dissemination of scientific data, is available on the Open Science Framework at Rodríguez-Prada et al. (2026a). This repository contains the scripts for data processing and analysis, which are also available at Rodríguez-Prada (2025).

Results

Analysis of Different Priors on Power, Type I Error Rate, and Model Selection in Bayesian Methods

Seven Bayesian priors were evaluated for estimating variance components across four population multilevel models under the simulated conditions (Table S1, Rodríguez-Prada et al., 2026b). Minimal differences were observed in statistical power among the priors ($F(6,\ 215136) =\ 0.76,\ p < .001, \eta_p^2 = 0.002$), Type I error rates ($F(6, 107568) = 0.34$, $p = 0.916$, $\eta_p^2 = < .001$), and model selection accuracy ($F(6, 428638) = 1.50$, $p = 0.174$; $\eta_p^2 = < .001$). Statistical power remained moderate (≈ 0.62), Type I error rates were conservative (≈ 0.033), and WAIC and LOO correctly identified the population model in 82% of cases. The truncated Cauchy prior (Half $-$ Cauchy $\sim (0, 10)$ was selected for further analyses due to slightly higher power and proximity to the nominal Type I error rate of 0.05, reflecting its suitability for this simulation context.

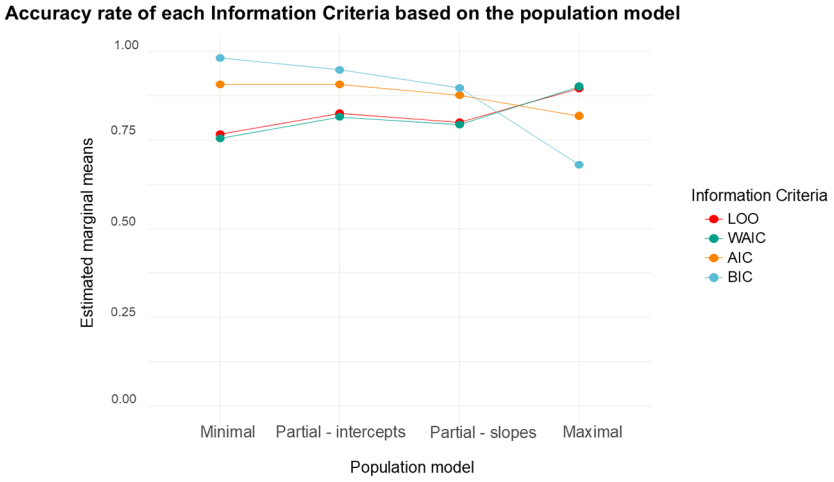
Model Selection Using Information Criteria

Correct model selection rates for frequentist (AIC, BIC) and Bayesian (WAIC, LOO) indices showed minimal overall differences ($F(1.64, 117212) = 1557.99$, $p < .001$; $\eta_p^2 = 0.007$). However, there was a significant interaction between the index and population model ($F(4.93, 117212.53) = 2883.85$, $p < .001$; $\eta_p^2 = .11$). Marginal means (Table S2, Rodríguez-Prada et al., 2026b) indicate that WAIC and LOO excel at identifying the maximal model but perform worse with simpler models, selecting them 76%–82% of the time. In contrast, frequentist indices favour simpler models: BIC overwhelmingly selects the minimal model (98%), while AIC demonstrates a more balanced across all population models (Figure 1). Although the interaction effects of the information criteria with both the number of individuals ($F(3.28, 117212.53) = 213.47$; $p < 0.001$; $\eta_p^2 = 0.002$) and repeated measures ($F(4.93, 117212.53) = 88.85$; $p < 0.001$; $\eta_p^2 = 0.001$) were small, they reveal noteworthy patterns. More individuals and repeated measures improve correct model selection. AIC and LOO are less sensitive to these factors, while frequentist indices, especially BIC, show greater variability. Overall, Figure 1 shows that BIC and AIC consistently outperform WAIC and LOO in simpler structures (minimal and partial–intercepts models), whereas their performance decreases when slope variability is included in the

model. In contrast, WAIC and LOO provide relatively stable, though lower, accuracy across models. Thus, the results suggest that the choice of information criterion has important implications: BIC and AIC tend to favor parsimony, whereas WAIC and LOO yield more balanced but less accurate selections.

Figure 1

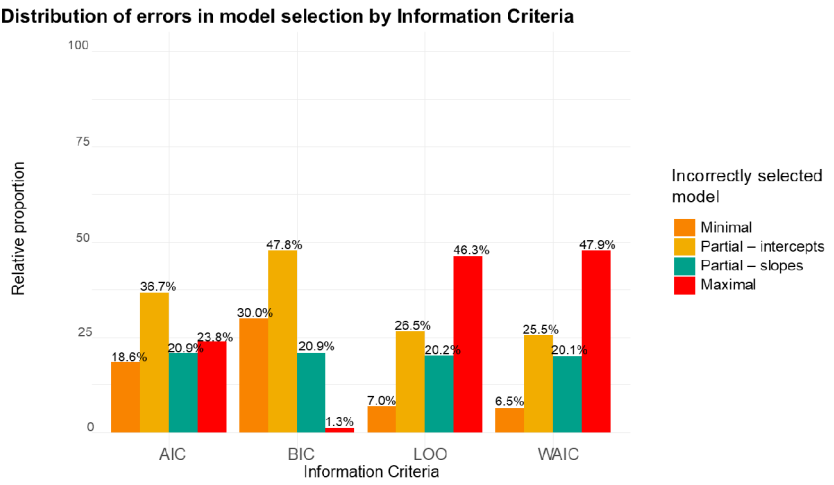
Proportion of Correct Selections (Accuracy Rate) for Each Relative Fit Index Based on Population Models



A detailed analysis of errors in BIC, AIC, WAIC, and LOO (i.e., the model selected when the fit index fails to identify the correct one) reveals distinct error patterns. BIC tends to overly penalise complex models and, when incorrect, selects one of the simpler models (minimal or partial-intercepts) in 77.8% of the occasions. In contrast, WAIC and LOO often err by selecting the maximal model nearly half the time. AIC shows the most balanced error distribution, although it slightly prefers the random-intercepts model (Figure 2). When AIC and BIC misidentify the true model, the most frequent error is selecting simpler models (e.g., BIC favours partial-intercepts at 47.8%). In contrast, WAIC and LOO tend to misclassify more complex models, with nearly half of errors involving the maximal specification (46.3% for LOO and 47.9% for WAIC). These results suggest that AIC and BIC are biased toward parsimony, whereas WAIC and LOO are more likely to overfit, highlighting systematic tendencies in model misclassification.

Figure 2

Distribution of Errors: Model Selection When the Information Criteria Fails to Identify the True Model



Power and Type I Error Rate Conditioned on Model Selection

Power and Type I error rates conditioned on model selection were analyzed, simulating scenarios in which the true population model is unknown, as happens in ecological contexts.

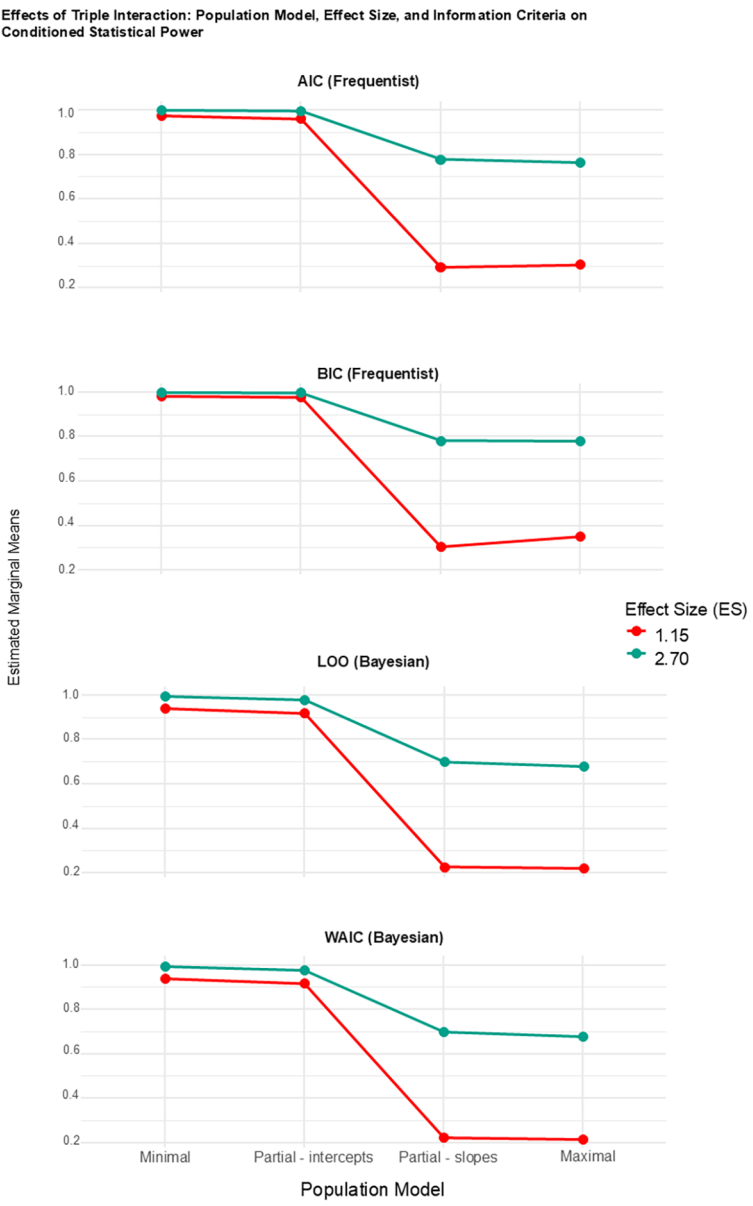
On power, ANOVA results showed a significant main effect of the fit index (F(1.27, 60132.31) = 2702.46; $p < 0.001$; $\eta_p^2 = 0.05$), with BIC yielding the highest power, followed by AIC, WAIC and LOO (Table S3, Rodríguez-Prada et al., 2026b). However, for BIC, this increase in power was accompanied by substantially inflated Type I error rates in more complex models. Significant effects were also observed for population model (F(3, 47482) = 11393.98; $p < 0.001$; $\eta_p^2 = 0.42$), effect size of the intervention (F(1, 47482) = 9713.90; $p < 0.001$; $\eta_p^2 = 0.17$), and number of individuals (F(2, 47482) = 2782.18; $p < 0.001$; $\eta_p^2 = 0.10$). Larger effects and sample sizes enhance power, whereas higher error variance in models, as in maximal models, diminishes it (Table S4, Rodríguez-Prada et al., 2026b).

The interaction between the population model and effect size (F(3, 47482) = 2309.91; $p < 0.001$; $\eta_p^2 = 0.108$) showed stable statistical power with large effect sizes across models. However, there were significant declines for moderate effects, particularly in Bayesian methods, reflecting their conservative nature (Figure 3; Table S4, Rodríguez-Prada et al., 2026b). Overall, across all criteria, statistical power is high and stable for simpler models (minimal and partial–intercepts) but declines sharply when slope variability and maximal structures are introduced. The drop is most pronounced for smaller effects (ES = 1.15), where power often falls below .40, particularly for LOO and WAIC. Larger effects (ES = 2.70) mitigate but do not eliminate this decline. These results indicate that model

complexity disproportionately reduces power for detecting smaller effects, with Bayesian indices (LOO, WAIC) being more sensitive to this loss than frequentist ones (AIC, BIC).

Figure 3

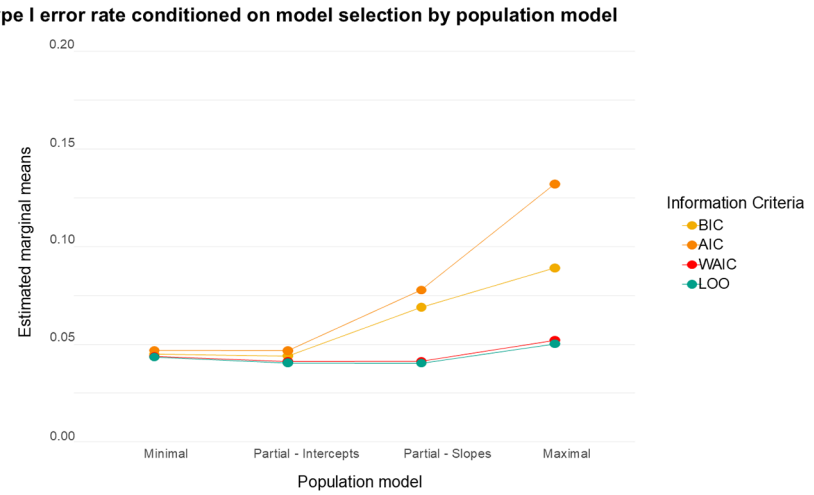
Effects of the Interaction Between Population Model, Effect Size and Fit Index on Conditioned Power



The model fit index, conditioned on the selected model, significantly impacts the Type I error rate ($F(1.49, 35636.95) = 511.52$, $p < .001$, $\eta_p^2 = .02$). Although the effect size is small, meaningful differences emerge among indices. Frequentist indices tend to exceed the nominal Type I error rate, while the Bayesian indices demonstrate a more conservative behaviour, staying closer to the nominal value of 0.05 (Table S5, Rodríguez-Prada et al., 2026b). The interaction between population model and model fit indices is statistically significant but small ($F(4.47, 35636.95) = 165.28$, $p < .001$, $\eta_p^2 = .02$). Examining marginal means (Table S5, Rodríguez-Prada et al., 2026b), the BIC index shows an unacceptable Type I error rate for complex models, reaching 13% for the maximal model. The AIC index is less sensitive than BIC, with a Type I error rate of 8.9% for the maximal model. WAIC and LOO remain stable across all population models, aligning closely with the nominal level in complex models (Figure 4). Thus, error rates remain close to the nominal .05 level for minimal and partial–intercepts models across all criteria. However, as complexity increases, AIC and BIC show inflated error rates, particularly under maximal specifications (AIC ≈ .13, BIC ≈ .09). By contrast, Bayesian fit indices (WAIC and LOO) maintain more stable control of Type I error, even in complex models. These findings suggest that AIC and BIC may increase the risk of false positives when applied to highly parameterized structures, whereas WAIC and LOO provide more conservative performance.

Figure 4

Type I Error Rate Conditioned on Model Selection as a Function of the Population Model



In summary, BIC exhibited the highest values for both statistical power and the Type I error rate. WAIC and LOO yielded the lowest values, remaining close to but not exceeding the nominal 0.05 level for false positives. AIC demonstrated a balanced performance between statistical power and the Type I error rate. Interaction effects highlighted the influence of between-subject factors on both power and the Type I error rate, particularly the intervention effect size, sample size, and population model for power; and only the interaction with the population model for the Type I error rate.

Discussion

Statistical modelling has historically been underutilised in SCEDs within clinical and psychological contexts due to inherent challenges including small sample sizes, limited observations, and reduced data availability. However, advances in methods such as MLMs and Bayesian approaches offer promising solutions (Baek & Ferron, 2013; Moeyaert et al., 2017). These methods are particularly valuable for estimating covariance parameters, enabling researchers to quantify between-individual subject variability and explore its causes (e.g., why some individuals benefit more from an intervention than others). At a descriptive level, multilevel models enable researchers to quantify intervention effects, estimate the magnitude and direction of changes in both levels and slopes, and assess their practical significance (Ferron et al., 2009). Several studies have focused on exploring estimation in terms of bias and standard errors (Baek et al., 2020; Ferron et al., 2010; Moeyaert et al., 2017). However, there has been limited focus on other statistical properties of MLMs. A key objective of this study was to compare the performance of model fit information criteria from two different frameworks (frequentist via REML estimation and Bayesian with varying levels of weakly-informativeness) in a context where the true population model is unknown.[1] Bayesian methods showed no significant differences in statistical power, Type I error rate, and model selection among the chosen priors, confirming previous studies (Baek et al., 2020; Moeyaert et al., 2017). The half-Cauchy and half-normal priors, which favor smaller variances, performed slightly better, as suggested in earlier research. However, prior selection is context-dependent, with effectiveness varying by intervention type, outcome characteristics, and research design. Commonly used priors, like half-normal or half-Cauchy distributions, are often based on simulations and may not suit empirical cases where the dependent variable has greater dispersion. Priors should be tailored to the outcome's scale and variability

1) It is worth noting that 'statistical significance' is not presented as evidence of population generalization nor of practical relevance in this study. We use it only as an operating property to calibrate model-selection procedures, acknowledging that random-effects (mis)specification can alter Type I error, power, and uncertainty estimates (e.g., confidence intervals, effect sizes, etc.). This perspective complements prior SCED work centered on bias and precision (Baek et al., 2020; Ferron et al., 2010; Moeyaert et al., 2017) and keeps the emphasis on effect sizes with uncertainty for practical interpretation.

for robust estimation. A stronger integration between simulation and applied studies is needed to improve these decisions. As highlighted by Moeyaert et al. (2017), prior calibration can be informed by reanalyses of empirical SCEDs which helps to ensure that the scale of the priors reflects realistic ranges for variance components. For applied researchers, however, translating prior knowledge into formal prior distributions remains challenging. More empirical work and meta-analytical evidence are needed to provide practical guidance on how to derive informative priors in SCED contexts.

The information criteria (AIC, BIC, WAIC, LOO) showed minor differences in their ability to identify the true data-generating model. This finding partially aligns with those of Moeyaert et al. (2017), which reported equivalence between REML and Bayesian methods for estimating fixed effects, as both frameworks produced similar results. However, interaction effects revealed nuanced behaviours: Bayesian indices tended to favour complex models, while frequentist indices were more accurate with simpler models. This pattern suggests that Bayesian methods may lean toward overparameterisation, whereas frequentist methods might prefer underparameterized models. More recent contributions have extended the scope of MLM research in SCEDs by exploring novel approaches, including the application of generalized linear models (GLMs) and refinements in variance component estimation. Li et al. (2024) recommends using AIC and BIC to select an optimal model in case of count data with overdispersion; but if there is overdispersion and zero-inflation, the recommendation is to use methods with lower penalty for these complex models. However, no Bayesian option was explored in Li et al. (2024). Recent simulation work has further emphasized the importance of model specification for variance components in SCEDs. Li et al. (2022) demonstrated that biased estimates of between-case variance are particularly problematic when the true variance is small, and that unconstrained optimization methods combined with post hoc model selection procedures (e.g., bootstrap-based RLRT) can improve estimation accuracy and inference. These findings highlight that not only the choice between frequentist and Bayesian approaches but also the technical details of variance component estimation and covariance structure specification critically affect the robustness of MLMs applied to SCEDs.

It is worth mentioning that there are some differences between fixed and random effects regarding ecological validity and generalization of models in SCEDs depending on their complexity. On the one hand, simpler models may have lower ecological validity despite their greater generalizability. At the same time, overparameterised models may lead to problems of model fit and reduced predictive validity, particularly when unnecessary fixed effects are added. In the field of SCEDs, including additional fixed effects can enrich the description of treatment effects, and adding more random effects is often crucial for adequately modelling change. But this increases the estimation challenges. Our findings align with previous literature suggesting that Bayesian methods may offer a valuable alternative to handle this kind of complex model (Baek et al., 2020; Moeyaert et al., 2017; Van den Noortgate & Onghena, 2003b). This may be particularly true consider-

ing that real SCEDs are generally more complex than those simulated in methodological studies, making the Bayesian framework potentially more suitable for real-world data. In contrast, frequentist approaches may be more appropriate for simpler scenarios as, for example, frequentist methods struggle to estimate covariance parameters in complex models (Moeyaert et al., 2017). AIC performs well across models, making it suitable for straightforward scenarios. In contrast, WAIC and LOO excel in complex models, being less affected by sample size and repeated measures, limitations often faced in SCEDs. BIC's performance varies and is less reliable with complex models. Each information criterion shows distinct error patterns: WAIC and LOO favor complex models, whereas AIC and BIC lean towards simpler ones, consistent with previous findings (Martínez-Huertas et al., 2022; Martínez-Huertas & Olmos 2022). This divergence may reflect frequentist methods' limitations in estimating random effects in complex settings, highlighting the importance of understanding the unique strengths of each index in different modelling contexts.

The analysis of power and Type I error rates was focused on emulating real-world conditions researchers face, where the true data-generation model is unknown. When the intervention effect size was large, no significant differences in statistical power were found between Bayesian and frequentist methods, indicating that either can offer reasonable conclusions. However, frequentist methods, especially BIC, exhibited higher Type I error rates in complex scenarios. Higher power driven by inflated Type I error is not desirable. AIC showed better performance with a moderate excess of Type I error (8.9%) and a good balance with statistical power. Bayesian methods, while slightly conservative (Error rate ≈ 0.04), offered more stable and consistent results across conditions. Therefore, despite limitations and considering all these findings, Bayesian methods may be more robust in complex SCED settings than frequentist methods.

Model specification significantly influences statistical performance as it affects the estimation of confidence intervals, standard errors, and effect sizes. For instance, when standardising mean differences, using biased standard errors can result in biased effect size estimates. Model detectability varies based on complexity, information criteria, and estimation methods. This study analysed four models of increasing complexity, ranging from one with no random effects to one with random intercepts and random slopes. In this simulation, simpler models, which assumed no individual variability in intervention effects (an unrealistic assumption), yield lower standard errors from residual variance, enhancing detectability but overlook stochastic dependencies in nested data, assuming independence (Moerbeek, 2004). While fitting well, they may not represent real scenarios. Additional variability from random slopes heightens standard errors of the intervention effect, complicating detection. Increased error variance in complex models might explain lower performance, with detectability improving only with larger intervention effect sizes. Bayesian methods excel in these scenarios, at least using weakly informative priors and fit indices like WAIC or LOO. In fact, models can become more complex with

relevant effects like temporal trends (Baek et al., 2020; Moeyaert et al., 2017), making them more intricate than those examined here. The MLMs simulated in this study were simpler, enabling controlled exploration of fit indices' performance. Unreported simulations revealed estimating the correlation between slopes and intercepts often yielded impossible values or failed. This simplicity may clarify the stronger performance of frequentist methods over Bayesian approaches in some conditions. However, as complexity increases, Bayesian methods are expected to outperform frequentist methods, especially with small sample sizes and few repeated measures.

Bayesian methods have distinct advantages for SCEDs, especially with complex random effects. They have been successfully applied in situations with heterogeneous Level-1 variances (Baek & Ferron, 2013) and autocorrelated error terms with trends (Natesan, 2019), which are challenging for frequentist methods. Bayesian approaches' flexibility makes them ideal for these complexities. Software like the *brms* package (Bürkner, 2017) enhances accessibility, enabling researchers to apply Bayesian methods without deep knowledge of computation. Other tools like JAGS, WinBUGS, STAN, or *rstan* further support users without advanced statistics expertise. The findings emphasise matching estimation methods to model complexity. AIC is suitable for simpler models due to its stability, while Bayesian methods are better for complex models or small samples. In any case, we recommend researchers to perform sensitivity analyses with multiple fit indices to ensure consistent model selection, enhancing their findings' robustness.

Limitations

The present simulation study evaluated the operating characteristics of Bayesian and frequentist methods in MLMs applied to SCEDs when the true population model is unknown. Findings are specific to simulated conditions and need empirical validation. Our findings should be interpreted conditional on the simulated data-generating mechanisms: an AB design with normally distributed outcomes and no explicit autocorrelation or phase-specific trends, nor cross-level interactions. In applied SCEDs, however, outcomes are often counts and time-related structure is common. These features can change the effective model complexity and the penalty-fit trade-off, potentially altering the relative behavior of AIC/BIC versus WAIC/LOO and frequentist versus Bayesian estimations. This is consistent with recent work extending multilevel SCED models to GLMM settings and more realistic covariance structures (e.g., Li et al., 2024, 2025a, 2025b), and motivates future simulations that jointly manipulate outcome distribution, trends, and autocorrelation to evaluate whether the performance patterns reported here generalize to these more ecological scenarios. These decisions limit the scope of our findings, since model complexity of MLM applications often arises precisely from elements such as trends (general or phase-specific) or autocorrelation (Natesan Batley & Hedges, 2021). We deliberately opted for a simple design to keep the conditions more controlled and to examine the performance of model selection criteria in a focused way, particularly

regarding random effects and the detection of random slopes. Including both baseline and intervention trends would have substantially increased the complexity of the simulation, making it more difficult to isolate the behaviour of the information. Future research should explore the behavior of information criteria on model selection in more complex designs (e.g., multiple baseline or reversal designs) and advanced covariance structures, such as autoregressive terms, to better reflect real-world data. Using dependent variables that follow normal distributions, uncommon in SCEDs where ceiling and floor effects and heteroscedasticity tend to feature, also limits generalizability. Future studies should investigate the behavior of both frequentist and Bayesian approaches within generalized linear multilevel models (GLMMs). By incorporating variables with asymmetric distributions, such as Poisson distributions, researchers can address this gap and explore small effect sizes that are frequently overlooked in qualitative analyses. Recent research has demonstrated that linear mixed models can be adapted to handle count outcomes in SCEDs, leading to more accurate estimates, particularly in cases of overdispersion or small sample sizes (Declercq et al., 2019). More recently, Li and colleagues have shown how GLMMs can be effectively applied to various count data scenarios, including zero-inflated and overdispersed distributions (Li, 2024; Li et al., 2025b; Li et al., 2024). They have also provided step-by-step tutorials designed for applied researchers (see Li et al., 2025a). These contributions emphasize the importance of moving beyond normally distributed outcomes to enhance the ecological validity and robustness of statistical inferences in SCED data.

Regarding the Bayesian fit indices, WAIC and LOO, both derived from the conditional likelihood, were chosen for their accessibility and compatibility with the evaluated models. However, future work could benefit from a broader exploration of Bayesian model comparison strategies. Greater emphasis should be placed on marginal rather than conditional likelihood-based approaches, which some authors argue are more suitable for model comparison in Bayesian frameworks (Ariyo et al., 2022; Merkle et al., 2019). Additionally, recent methodological advances demonstrate the growing applicability of Bayes factors within different SCEDs, such as ABAB, alternating treatments and changing criterion designs (Yamada & Okada, 2024, 2025). Incorporating Bayes factors in future analyses could offer complementary insights into model selection decisions. The best results of these information criteria were observed under optimal conditions: more individuals, repeated measures, and larger effect sizes, consistent with previous research (Baek et al., 2020; Moeyaert et al., 2017). While MLMs effectively capture complex realities, their mathematical and statistical demands underscore the importance of clear model specification. Bickel (2007) observes that these models perform best when grounded in robust theories and literature. Thus, advancing MLMs should also promote theoretical development, particularly within psychology. It is important to remember that statistical decisions should never be made without considering the wider context and the insights provided by clinical expertise. An analysed effect might not reach the

traditional thresholds of statistical significance but can still be socially relevant, aligning with the concept of social validity (Kazdin, 1977; Snodgrass et al., 2023), and ways of control the antecession of the intervention before the change is fundamental to reassure the efficacy of an intervention (Perone, 1999). Future research should investigate the contingency relations between the decisions proposed by these models and those made by applied researchers, to understand how they are related.

Conclusions and Recommendations

MLM is a particularly versatile option for analysing SCED data, offering general and individual effect quantification, statistical testing, and the ability to model complex effects. This study highlights the strengths of Bayesian methods for highly complex models, where they maintain stable power and Type I error rates across varying effect sizes, sample sizes, and repeated measures. For simpler models, the frequentist REML approach performs equally well or better, particularly when AIC is used for model selection. When protection against Type I errors is paramount, as in assessing therapeutic change, Bayesian methods provide a reliable and robust framework, making them a valuable tool for applied researchers in SCEDs. Given that the data-generating model is typically unknown in applied SCEDs, the practical contribution of this study is to characterise the pros and cons of common model selection fit indices. Under our simulation conditions, REML estimations with AIC offered the most favourable balance between power and Type I error rates in simpler population models. But Bayesian estimations provided a more stable positive performance when model complexity increased. Thus, our recommendation is to treat model selection as risk management under uncertainty, triangulating across plausible random-effects structures, and using sensitivity analyses across AIC/BIC and WAIC/LOO to ensure that conclusions are not criterion-dependent.

Funding: CRP was supported by the "Ayudas al Fomento de la Investigación en Másteres Oficiales 2019-2020" and "Ayudas al Fomento de la Investigación en Másteres Oficiales 2020-2021" by the Universidad Autónoma de Madrid. CRP is also supported by "Contratos predoctorales para la Formación de Personal Investigador FPI-UAM 2022", Universidad Autónoma de Madrid, Spain.

Acknowledgments: Thanks to the Centro de Computación Científica (https://www.ccc.uam.es/) of the Universidad Autónoma de Madrid for the resources they provided for the simulation process.

Competing Interests: The authors have declared that no competing interests exist.

Author Contributions: *Cristina Rodríguez-Prada*: Conceptualization, Data Curation, Methodology, Formal Analysis, Software, Visualization, Writing – Original draft, Writing – Review & editing. *José Ángel Martínez-Huertas*: Conceptualization, Methodology, Writing – Review & editing. *Ricardo Olmos*: Conceptualization, Data Curation, Methodology, Supervision, Software, Writing – Review & editing.

Data Availability: The study dataset, code scripts for data, and supplementary materials that support the findings of this study are available in the OSF repository at Rodríguez-Prada et al. (2026a). Supplementary tables for this study are available at Rodríguez-Prada et al. (2026b).

Supplementary Materials

Type of supplementary materials	Availability/Access
Data	
Study dataset.	Rodríguez-Prada et al. (2026a)
Code	
R code scripts for data.	Rodríguez-Prada et al. (2026a)
Material	
Supplementary materials.	Rodríguez-Prada et al. (2026a)
Supplementary tables.	Rodríguez-Prada et al. (2026b)
Study/Analysis preregistration	
The study was not preregistered.	—
Other	
No other materials available.	—

References

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In E. Parzen, K. Tanabe & G. Kitagawa (Eds.), *Selected papers of Hirotugu Akaike* (pp. 199–213). Springer. https://doi.org/10.1007/978-1-4612-1694-0_15

Ariyo, O., Lesaffre, E., Verbeke, G., & Quintero, A. (2022). Model selection for Bayesian linear mixed models with longitudinal data: Sensitivity to the choice of priors. *Communications in Statistics — Simulation and Computation, 51*(4), 1591–1615. https://doi.org/10.1080/03610918.2019.1676439

Baek, E., Beretvas, S. N., Van den Noortgate, W., & Ferron, J. M. (2020). Brief research report: Bayesian versus REML estimations with noninformative priors in multilevel single-case data. *Journal of Experimental Education, 88*(4), 698–710. https://doi.org/10.1080/00220973.2018.1527280

Baek, E. K., & Ferron, J. M. (2013). Multilevel models for multiple-baseline data: Modeling across-participant variation in autocorrelation and residual variance. *Behavior Research Methods, 45*(1), 65–74. https://doi.org/10.3758/s13428-012-0231-z

Baek, E., & Ferron, J. M. (2020). Modeling heterogeneity of the Level-1 error covariance matrix in multilevel models for single-case data. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 16*(2), 166–185. https://doi.org/10.5964/meth.2817

Bickel, R. (2007). *Multilevel analysis for applied research: It's just regression!* Guilford Press.

Bono, R., & Arnau, J. (2014). *Diseños de caso único en ciencias sociales y de la salud [Unique case designs in social sciences and health sciences]*. Síntesis.

Botella, J., & Caperos, J. M. (2019). *Metodología de investigación en psicología general sanitaria [Research methodology in general health psychology]*. Síntesis.

Brooks, S. (1998). Markov chain Monte Carlo method and its application. *Journal of the Royal Statistical Society: Series D, 47*(1), 69–100. https://doi.org/10.1111/1467-9884.00117

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Busk, P. L., & Serlin, R. C. (1992). Meta-analysis for single-case research. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 187–212). Lawrence Erlbaum Associates.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge.

Declercq, L., Jamshidi, L., Fernández-Castilla, B., Beretvas, S. N., Moeyaert, M., Ferron, J. M., & Van den Noortgate, W. (2019). Analysis of single-case experimental count data using the linear mixed effects model: A simulation study. *Behavior Research Methods, 51*(6), 2477–2497. https://doi.org/10.3758/s13428-018-1091-y

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology, Research and Practice, 40*(5), 532–538. https://doi.org/10.1037/a0015808

Ferron, J. M., Bell, B. A., Hess, M. R., Rendina-Gobioff, G., & Hibbard, S. T. (2009). Making treatment effect inferences from multiple-baseline data: The utility of multilevel modeling approaches. *Behavior Research Methods, 41*(2), 372–384. https://doi.org/10.3758/BRM.41.2.372

Ferron, J. M., Farmer, J. L., & Owens, C. M. (2010). Estimating individual treatment effects from multiple-baseline data: A Monte Carlo study of multilevel-modeling approaches. *Behavior Research Methods, 42*(4), 930–943. https://doi.org/10.3758/BRM.42.4.930

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (Comment on article by Browne and Draper). *Bayesian Analysis, 1*(3), 515–534. https://doi.org/10.1214/06-BA117A

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). CRC Press.

Gentile, J. R., Roden, A. H., & Klein, R. D. (1972). An analysis-of-variance model for the intrasubject replication design. *Journal of Applied Behavior Analysis, 5*(2), 193–198. https://doi.org/10.1901/jaba.1972.5-193

Hoffman, L. (2014). *Longitudinal analysis: Modeling within-person fluctuation and change* (1st ed.). Routledge.

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research, 15*, 1593–1623.

Kazdin, A. E. (1977). Assessing the clinical or applied importance of behavior change through social validation. *Behavior Modification, 1*(4), 427–452. https://doi.org/10.1177/014544557714001

Kazdin, A. E. (1982). *Single-case research designs: Methods for clinical and applied settings*. Oxford University Press.

Keselman, H. J., & Leventhal, L. (1974). Concerning the statistical procedures enumerated by Gentile et al.: Another perspective. *Journal of Applied Behavior Analysis, 7*(4), 643–645. https://doi.org/10.1901/jaba.1974.7-643

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2013). Single-case intervention research design standards. *Remedial and Special Education, 34*(1), 26–38. https://doi.org/10.1177/0741932512452794

Li, H. (2024). Model selection of GLMMs in the analysis of count data in single-case studies: A Monte Carlo simulation. *Behavior Research Methods, 56*(7), 7963–7984. https://doi.org/10.3758/s13428-024-02464-7

Li, H., Baek, E., Luo, W., Du, W., & Lam, K. H. (2025a). Using generalized linear mixed models in the analysis of count and rate data in single-case experimental designs: A step-by-step tutorial. *Evaluation & the Health Professions, 48*(1), 143–155. https://doi.org/10.1177/01632787241259500

Li, H., Luo, W., & Baek, E. (2024). Multilevel modeling in single-case studies with zero-inflated and overdispersed count data. *Behavior Research Methods, 56*(4), 2765–2781. https://doi.org/10.3758/s13428-024-02359-7

Li, H., Luo, W., Baek, E., Thompson, C. G., & Lam, K. H. (2022). Estimation and statistical inferences of variance components in the analysis of single-case experimental design using multilevel modeling. *Behavior Research Methods, 54*(4), 1559–1579. https://doi.org/10.3758/s13428-021-01691-6

Li, H., Luo, W., Baek, E., Thompson, C. G., & Lam, K. H. (2025b). Multilevel modeling in single-case studies with count and proportion data: A demonstration and evaluation. *Psychological Methods, 30*(4), 815–842. https://doi.org/10.1037/met0000607

Manolov, R., Sierra, V., Solanas, A., & Botella, J. (2014). Assessing functional relations in single-case designs: Quantitative proposals in the context of the evidence-based movement. *Behavior Modification, 38*(6), 878–913. https://doi.org/10.1177/0145445514545679

Manolov, R., & Moeyaert, M. (2017). Recommendations for choosing single-case data analytical techniques. *Behavior Therapy, 48*(1), 97–114. https://doi.org/10.1016/j.beth.2016.04.008

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. https://doi.org/10.1016/j.jml.2017.01.001

Martínez-Huertas, J. A., & Olmos, R. (2022). Recovering crossed random effects in mixed-effects models using model averaging. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 18*(4), 298–323. https://doi.org/10.5964/meth.9597

Martínez-Huertas, J. Á., Olmos, R., & Ferrer, E. (2022). Model selection and model averaging for mixed-effects models with crossed random effects for subjects and items. *Multivariate Behavioral Research, 57*(4), 603–619. https://doi.org/10.1080/00273171.2021.1889946

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan* (2nd ed.). Chapman and Hall/CRC.

McNeish, D. (2016). On using Bayesian methods to address small sample problems. *Structural Equation Modeling, 23*(5), 750–773. https://doi.org/10.1080/10705511.2016.1186549

Merkle, E. C., Furr, D., & Rabe-Hesketh, S. (2019). Bayesian comparison of latent variable models: Conditional versus marginal likelihoods. *Psychometrika, 84*(3), 802–829. https://doi.org/10.1007/s11336-019-09679-0

Moerbeek, M. (2004). The consequence of ignoring a level of nesting in multilevel analysis. *Multivariate Behavioral Research, 39*(1), 129–149. https://doi.org/10.1207/s15327906mbr3901_5

Moeyaert, M., Ferron, J. M., Beretvas, S. N., & Van den Noortgate, W. (2014). From a single-level analysis to a multilevel analysis of single-case experimental designs. *Journal of School Psychology, 52*(2), 191–211. https://doi.org/10.1016/j.jsp.2013.11.003

Moeyaert, M., Manolov, R., & Rodabaugh, E. (2020). Meta-analysis of single-case research via multilevel models: Fundamental concepts and methodological considerations. *Behavior Modification, 44*(2), 265–295. https://doi.org/10.1177/0145445518806867

Moeyaert, M., Rindskopf, D., Onghena, P., & Van den Noortgate, W. (2017). Multilevel modeling of single-case data: A comparison of maximum likelihood and Bayesian estimation. *Psychological Methods, 22*(4), 760–778. https://doi.org/10.1037/met0000136

Moeyaert, M., & Yang, P. (2021). Assessing generalizability and variability of single-case design effect sizes using two-stage multilevel modeling including moderators. *Behaviormetrika, 48*(2), 207–229. https://doi.org/10.1007/s41237-021-00141-z

Moeyaert, M., Yang, P., & Xue, Y. (2024). Individual participant data meta-analysis including moderators: Empirical validation. *Journal of Experimental Education, 92*(4), 723–740. https://doi.org/10.1080/00220973.2023.2208062

Natesan, P. (2019). Fitting Bayesian models for single-case experimental designs. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 15*(4), 147–156. https://doi.org/10.1027/1614-2241/a000180

Natesan Batley, P., & Hedges, L. V. (2021). Accurate models vs. accurate estimates: A simulation study of Bayesian single-case experimental designs. *Behavior Research Methods, 53*(4), 1782–1798. https://doi.org/10.3758/s13428-020-01522-0

Nicenboim, B., & Vasishth, S. (2016). Statistical methods for linguistic research: Foundational ideas — Part II. *Language and Linguistics Compass, 10*(11), 591–613. https://doi.org/10.1111/lnc3.12207

Parker, R. I., Vannest, K. J., & Davis, J. L. (2011). Effect size in single-case research: A review of nine nonoverlap techniques. *Behavior Modification, 35*(4), 303–322. https://doi.org/10.1177/0145445511399147

Parsonson, B. S., & Baer, D. M. (1986). The graphic analysis of data. In A. Poling & R. W. Fuqua (Eds.), *Research methods in applied behavior analysis: Issues and advances* (pp. 157–186). Springer US. https://doi.org/10.1007/978-1-4684-8786-2_8

Perone, M. (1999). Statistical inference in behavior analysis: Experimental control is better. *Behavior Analyst, 22*(2), 109–116. https://doi.org/10.1007/BF03391988

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2021). *nlme: Linear and nonlinear mixed effects models* (Version 3.1-152) [Software]. https://CRAN.R-project.org/package=nlme

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25*, 111–163. https://doi.org/10.2307/271063

Richardson, J. T. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review, 6*(2), 135–147. https://doi.org/10.1016/j.edurev.2010.12.001

Rindskopf, D. (2014). Nonlinear Bayesian analysis for single case designs. *Journal of School Psychology, 52*(2), 179–189. https://doi.org/10.1016/j.jsp.2013.12.003

Rodabaugh, E., & Moeyaert, M. (2017). Multilevel modeling of single-case data: An introduction and tutorial for the applied researcher. *NERA Conference Proceedings 2017(8), University of Connecticut*. https://opencommons.uconn.edu/nera-2017/8

Rodríguez-Prada, C. (2025). *Bayesian versus frequentist in SCEDs — R code — v1.0.0* [Github Project page containing R code for study simulation]. GitHub. https://github.com/Cristrinaranjus/phdthesis/releases/tag/methodology-journal

Rodríguez-Prada, C., & Olmos, R. (2019). Análisis multinivel y medidas del tamaño del efecto en diseños de caso único: Un estudio piloto comparando datos de simulación y datos empíricos [Multilevel analysis and effect size measures in single-case designs: A pilot study comparing simulated and empirical data]. *VIII Congreso SAVECC - Sociedad para el Avance del Estudio Científico del Comportamiento[8th SAVECC Congress: Society for the Advancement of the Scientific Study of Behavior)]*. Universidad Autónoma de Madrid. https://doi.org/10.13140/RG.2.2.33619.45609

Rodríguez-Prada, C., Olmos, R., & Martínez-Huertas, J. Á. (2026a). *Bayesian versus frequentist approaches in multilevel single-case designs: On Type I error rate and power* [OSF project page

containing study dataset, code scripts for data, supplementary materials]. Open Science Foundation. https://doi.org/10.17605/OSF.IO/K7B82

Rodríguez-Prada, C., Martínez-Huertas, J. Á., & Olmos, R. (2026b). *Supplementary materials to* "Bayesian versus frequentist approaches in multilevel single-case designs: On Type I error rate and power" [Supplementary tables]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.21777

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods, 43*(4), 971–980. https://doi.org/10.3758/s13428-011-0111-y

Snodgrass, M., Cook, B. G., & Cook, L. (2023). Considering social validity in special education research. *Learning Disabilities Research & Practice, 38*(4), 311–319. https://doi.org/10.1111/ldrp.12326

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. https://doi.org/10.1111/1467-9868.00353

Van de Schoot, R., Broere, J. J., Perryck, K. H., Zondervan-Zwijnenburg, M., & van Loey, N. E. (2015). Analyzing small data sets using Bayesian estimation: The case of posttraumatic stress symptoms following mechanical ventilation in burn survivors. *European Journal of Psychotraumatology, 6*, Article 25216. https://doi.org/10.3402/ejpt.v6.25216

Van den Noortgate, W., & Onghena, P. (2003a). Combining single-case experimental data using hierarchical linear models. *School Psychology Quarterly, 18*(3), 325–346. https://doi.org/10.1521/scpq.18.3.325.22577

Van den Noortgate, W., & Onghena, P. (2003b). Hierarchical linear models for the quantitative integration of effect sizes in single-case research. *Behavior Research Methods, Instruments, & Computers, 35*(1), 1–10. https://doi.org/10.3758/BF03195492

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing, 27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Virués-Ortega, J., & Haynes, S. N. (2005). Functional analysis in behavior therapy: Behavioral foundations and clinical application. *International Journal of Clinical and Health Psychology, 5*(3), 567–587.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research, 11*(116), 3571–3594.

Weakliem, D. L. (1999). A critique of the Bayesian information criterion for model selection. *Sociological Methods & Research, 27*(3), 359–397. https://doi.org/10.1177/0049124199027003002

Yamada, T., & Okada, K. (2024). Bayes factor for single-case ABAB design data. *Behaviormetrika, 51*(1), 277–286. https://doi.org/10.1007/s41237-023-00206-1

Yamada, T., & Okada, K. (2025). Bayes factor for major single-case experimental designs: Case for alternating treatment design and changing criterion design. *Behaviormetrika, 52*, 707–720. https://doi.org/10.1007/s41237-025-00259-4