

A randomization test wrapper for synthesizing single-case experiments using multilevel models: A Monte Carlo simulation study

Bart Michiels¹ · René Tanious¹ · Tamal Kumar De¹ · Patrick Onghena¹

Published online: 3 July 2019
© The Psychonomic Society, Inc. 2019

Abstract

Multilevel models (MLMs) have been proposed in single-case research, to synthesize data from a group of cases in a multiple-baseline design (MBD). A limitation of this approach is that MLMs require several statistical assumptions that are often violated in single-case research. In this article we propose a solution to this limitation by presenting a randomization test (RT) wrapper for MLMs that offers a nonparametric way to evaluate treatment effects, without making distributional assumptions or an assumption of random sampling. We present the rationale underlying the proposed technique and validate its performance (with respect to Type I error rate and power) as compared to parametric statistical inference in MLMs, in the context of evaluating the average treatment effect across cases in an MBD. We performed a simulation study that manipulated the numbers of cases and of observations per case in a dataset, the data variability between cases, the distributional characteristics of the data, the level of autocorrelation, and the size of the treatment effect in the data. The results showed that the power of the RT wrapper is superior to the power of parametric tests based on *F* distributions for MBDs with fewer than five cases, and that the Type I error rate of the RT wrapper is controlled for bimodal data, whereas this is not the case for traditional MLMs.

Keywords Multiple-baseline design · Multilevel model · Randomization test · Power analysis · Monte Carlo simulation study

Multilevel models¹ (MLMs) are frequently used to analyze nested data in various subfields of the behavioral and the social sciences. Examples of this type of data include repeated measurements of individuals in longitudinal research, students that are nested in schools in educational research, or employees that are nested in companies in organizational psychology. MLMs have also been proposed for the statistical analysis of single-case experimental designs (SCEDs; e.g., Ferron, Bell, Hess, Rendina-Gobioff, & Hibbar, 2009;

Ferron, Farmer, & Owens, 2010; Jenson, Clark, Kircher, & Kristjánsson, 2007; Van den Noortgate & Onghena, 2003a, 2003b, 2007, 2008). SCEDs are a group of experimental designs that are increasingly being used in various fields of the behavioral sciences, including special education (Alnahdi, 2015), school psychology (Swaminathan & Rogers, 2007), and clinical psychology (Kazdin, 2011), as well as in the medical sciences, where they are referred to as “*N*-of-1 designs” (Gabler, Duan, Vohra, & Kravitz, 2011). In contrast to case studies or other nonexperimental research, SCEDs are designed experiments in which a single entity is measured over time on one or more dependent variables under different levels (i.e., treatments) of one or more independent variables (Barlow, Nock, & Hersen, 2009). Note that “entity” can refer to various units, such as a single person, a classroom, or a group of subjects (Levin, O’Donnell, & Kratochwill, 2003).

One of the simplest SCEDs is the AB phase design, in which a single subject is measured repeatedly during a baseline phase (A phase) and a subsequent treatment phase (B phase). AB phase designs are often replicated across different persons, behaviors, or settings to evaluate the generalizability of the results (for convenience, we will refer to all of these types of replications as “cases”). AB phase designs can be

¹ Multilevel models are known by different names, depending on the discipline (Raudenbush & Bryk, 2002). Some synonyms include “hierarchical regression models,” “mixed effects models,” “random-coefficient regression models,” and “covariance component models.” In this article, we will use the term “multilevel models.”

Electronic supplementary material The online version of this article (<https://doi.org/10.3758/s13428-019-01266-6>) contains supplementary material, which is available to authorized users.

✉ Patrick Onghena
patrick.onghena@kuleuven.be

¹ Faculty of Psychology and Educational Sciences, KU Leuven, Leuven, Belgium

replicated in either sequential replication designs or simultaneous replication designs (Onghena, 2005). In *sequential* replication designs, multiple SCEDs are executed one after another, whereas in *simultaneous* replication designs, multiple SCEDs are executed simultaneously. The multiple-baseline design (MBD) across participants is a simultaneous replication design that consists of several single-case AB phase designs (Onghena & Edgington, 2005). A survey by Shadish and Sullivan (2011) that investigated the characteristics of a large body of published single-case research showed that more than half of the surveyed studies utilized an MBD across participants, indicating that these designs are used very often in single-case research.

An advantage for the internal validity of an MBD over sequentially replicated AB phase designs is that the data for the participants are collected concurrently, which enables between-series comparisons, in addition to the within-series AB comparisons that are also possible in sequentially replicated AB phase designs. More specifically, the intervention can be introduced in a time-staggered way, so that the length of each case's A phase is extended a little further than the previous case's. In this way, researchers can compare cases already in the treatment phase to cases that are still in the baseline phase at the same point in time. The possibility of such between-series comparisons strengthens the internal validity of the MBD because any observed effects may be more confidently attributed to the introduction of the treatment rather than to external events that might affect all cases (Baer, Wolf, & Risley, 1968; Kazdin, 2011; Koehler & Levin, 2000). Furthermore, it is generally recommended to randomize the time-staggered intervention points in an MBD, because this further increases the internal validity of the design (Edgington, 1969, 1996; Heyvaert, Wendt, Van den Noortgate, & Onghena, 2015; Levin, Ferron, & Gafurov, 2018; Kratochwill & Levin, 2010; Marascuilo & Busk, 1988; Tyrrell, Corey, Feldman, & Silverman, 2013; Wampold & Worsham, 1986). More specifically, randomizing the intervention points can statistically control for confounding variables that are time-related.

In MBDs the repeated measurements are nested within cases, and as such, the entire dataset can be modeled by a two-level model that allows estimation of the average treatment effect across cases as well as case-specific treatment effects (Van den Noortgate & Onghena, 2003a, 2003b). Furthermore Onghena, Michiels, Jamshidi, Moeyaert, and Van den Noortgate (2018) noted that MLMs constitute a versatile and comprehensive framework for the analysis and meta-analysis of SCEDs that connects to general statistical theory. A two-level model is defined by the following regression equation at the first level:

$$y_{ij} = \beta_{0j} + \beta_{1j} * Phase_{ij} + e_{ij}, \text{ and } e_{ij} \sim N(0, \sigma_e^2), \quad (1)$$

with y_{ij} being the value of the outcome variable for case j at measurement occasion i , β_{0j} the regression intercept (i.e., the mean level of the baseline phase) for case j , β_{1j} the regression coefficient of the treatment effect for case j , $Phase_{ij}$ a dummy variable that takes the value of 0 in the baseline phase and the value of 1 in the treatment phase, and e_{ij} the residuals of the model, which are assumed to be normally distributed with a mean of 0 and a variance of σ_e^2 . For the second level, the following regression equations can be used:

$$\begin{cases} \beta_{0j} = \theta_{00} + u_{0j} \\ \beta_{1j} = \theta_{10} + u_{1j} \end{cases}, \text{ and } \begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{u0}^2 & \sigma_{u0u1} \\ \sigma_{u1u0} & \sigma_{u1}^2 \end{pmatrix} \right). \quad (2)$$

At the second level of the MLM, the case-specific intercepts (β_{0j}) and regression coefficients (β_{1j}) are modeled using separate parameters for the average baseline level (θ_{00}) and treatment effect (θ_{10}) across participants and for the case-specific deviations from these average values (u_{0j} and u_{1j} , respectively). When applied to modeling MBD data, the inclusion of case-specific residual error terms (u_{0j} and u_{1j}) makes sense because it is unlikely that baseline levels and treatment effect sizes are identical across all cases. Furthermore, the model assumes that u_{0j} and u_{1j} are multivariate normally distributed with means of 0, variances of σ_{u0}^2 and σ_{u1}^2 , respectively, and a covariance of $\sigma_{u1u0} = \sigma_{u0u1}$. Although all parameters within the two-level model can be of potential interest and relevance, single-case researchers are usually mainly interested in the average treatment effect over the included cases (θ_{10}).²

The variance–covariance matrix of MLMs is traditionally estimated using maximum likelihood procedures (ML; Raudenbush & Bryk, 2002). An advantage of ML estimates is that they are consistent and asymptotically normal, which enables parametric significance testing and the construction of confidence intervals. MLMs can be estimated using either full maximum likelihood (FML) or restricted maximum likelihood (REML). The difference between FML and REML is that in REML the variances and covariances are estimated after controlling for fixed effects, which results in less biased variance and covariance components (Harville, 1977). Given that ML procedures are based on large-sample theory, it follows that sample sizes must be large in order for ML estimates to be unbiased. Although sample size recommendations for Level-1 and Level-2 units of MLMs may vary considerably (e.g., Clarke & Wheaton, 2007; Maas & Hox, 2004), it is clear that in the context of analyzing single-case MBDs, sample sizes are far too small for the asymptotic properties of ML estimates to apply. Consequently, statistical inferences about

² Equations 1 and 2 only represent the most basic two-level model. Three- and higher-level models are possible, and more complex error structures can also be included, to model, for example, autocorrelated errors (see, e.g., Onghena et al., 2018).

ML estimates cannot rely on these properties. A simulation study by Ferron et al. (2009) showed that although estimates for fixed effects are still unbiased in the context of analyzing single-case research, the estimates of the variance components are biased. This bias in variance components results in biased standard errors for the fixed effects, and thus inflated or deflated Type I error rates for statistical inferences about fixed effects based on t or F tests (Moeyaert, Ugille, Ferron, Beretvas, & Van den Noortgate, 2013). Adjusted F -test procedures have been proposed in an attempt to remedy this issue. For example, the Satterthwaite approximation involves adjusting the estimated degrees of freedom of the variance–covariance matrix so that the F ratio is approximately correct (Fai & Cornelius, 1996; Satterthwaite, 1941). The Kenward–Roger approximation is an extension of the Satterthwaite approximation and includes an adjustment for small-sample size bias, which involves adjusting both the F statistic and the degrees of freedom and results in more conservative p values than the latter approximation (Kenward & Roger, 1997). However, a problem with these adjustment procedures is that they do not deal with the uncertainty regarding the variance component estimates for small sample sizes (Burrick & Graybill, 1992; Kenward & Roger, 2009).

Another way to estimate the variance–covariance matrix of an MLM is by Bayesian estimation (Baldwin & Fellingham, 2013; Browne, Draper, Goldstein, & Rasbash, 2002; Shadish et al., 2014). Moeyaert, Rindskopf, Onghena, and Van den Noortgate (2017) compared the bias of the fixed effects and variance components for both ML estimation and Bayesian estimation procedures (using weakly informative priors; priors that are intentionally weaker than the actual evidence that is available). With respect to the estimation of fixed effects, the results showed that ML estimation and Bayesian estimation produced very similar estimates. However, both the ML and Bayesian variance estimates were biased and imprecisely estimated if there were only three participants. When the number of participants was increased to seven, the relative bias was close to 5% and the estimates were more precise. In addition, when priors were more informative in the Bayesian estimation procedure, both the fixed effects and the variance components could be estimated more precisely. Although the use of informative priors yielded the most precise results, Moeyaert et al. (2017) argued that the use of weakly informative priors is the most appropriate choice for single-case research. One justification of this argument is that the choice of a prior distribution can have a substantial impact on the resulting statistical inferences of the MLM, especially given the small sample sizes that are common in single-case research (Gelman, 2006; Gelman, Carlin, Stern, & Rubin, 2013). A second reason is that weakly informative priors still allow the data to speak for themselves, without the prior having too large an influence on the overall model (Spiegelhalter, Abrams, & Myles, 2004).

In this article, we propose a nonparametric approach for making statistical inferences regarding the fixed effects of an MLM, by using MLMs within a randomization test (RT) framework. In this approach, fixed-effect estimates can be used as the test statistic in an RT, and nonparametric p values can be derived for these estimates without requiring distributional assumptions such as normality or an assumption of random sampling, and without being dependent on potentially biased variance components of the MLM. From the outset, we want to emphasize that the validity of this approach is dependent on the requirement that some form of experimental randomization be present in the design. As such, this approach is only valid for randomized single-case designs.

In the following paragraphs we will first use empirical data to demonstrate how an MBD can be analyzed with respect to the average treatment effect across cases, in a two-level model with REML estimation and Kenward–Roger-adjusted F tests. Second, we will introduce the RT, explain the rationale of the test, and illustrate its use with the same dataset. Third, we will explain how both the MLM and RT procedures can be combined in order to enable nonparametric inferences for the average treatment effect in an MLM that have guaranteed nominal Type I error rate control. Fourth, we will compare this combined MLM-RT and MLM with respect to Type I error rate and power for MBDs with small numbers of participants and measurement occasions, by means of a Monte Carlo simulation study.

Analyzing multiple-baseline designs with multilevel models using maximum likelihood estimation

In this section, we illustrate the use of a two-level model to test the null hypothesis of no average treatment effect on data collected by Franco, Davis, and Davis (2013). These authors used an MBD across six nonverbal children with autism who were taught to engage in social interaction within play routines. The children's social interaction was assessed by using behavioral observation techniques and video coding of all intentional communication acts, defined as any attempt that the child made to interact with an adult within the social routine (using vocalizations, gestures, or eye gaze). The main outcome variable consisted of the maximum number of child actions to maintain social interaction during a single social routine.

The data for this illustration were recovered from Fig. 1 of Franco et al. (2013) using the “GetData Graph Digitizer,” version 2.26 (Fedorov, 2013). The study by Franco et al. had unequal numbers of measurements for each case (17, 18, 19, 20, 22, and 24) and included two follow-up sessions, but for the following illustrations, we limited the analyses to the A and B phases for which complete data for all cases were available (i.e., the first 17 sessions of the six children). We added this

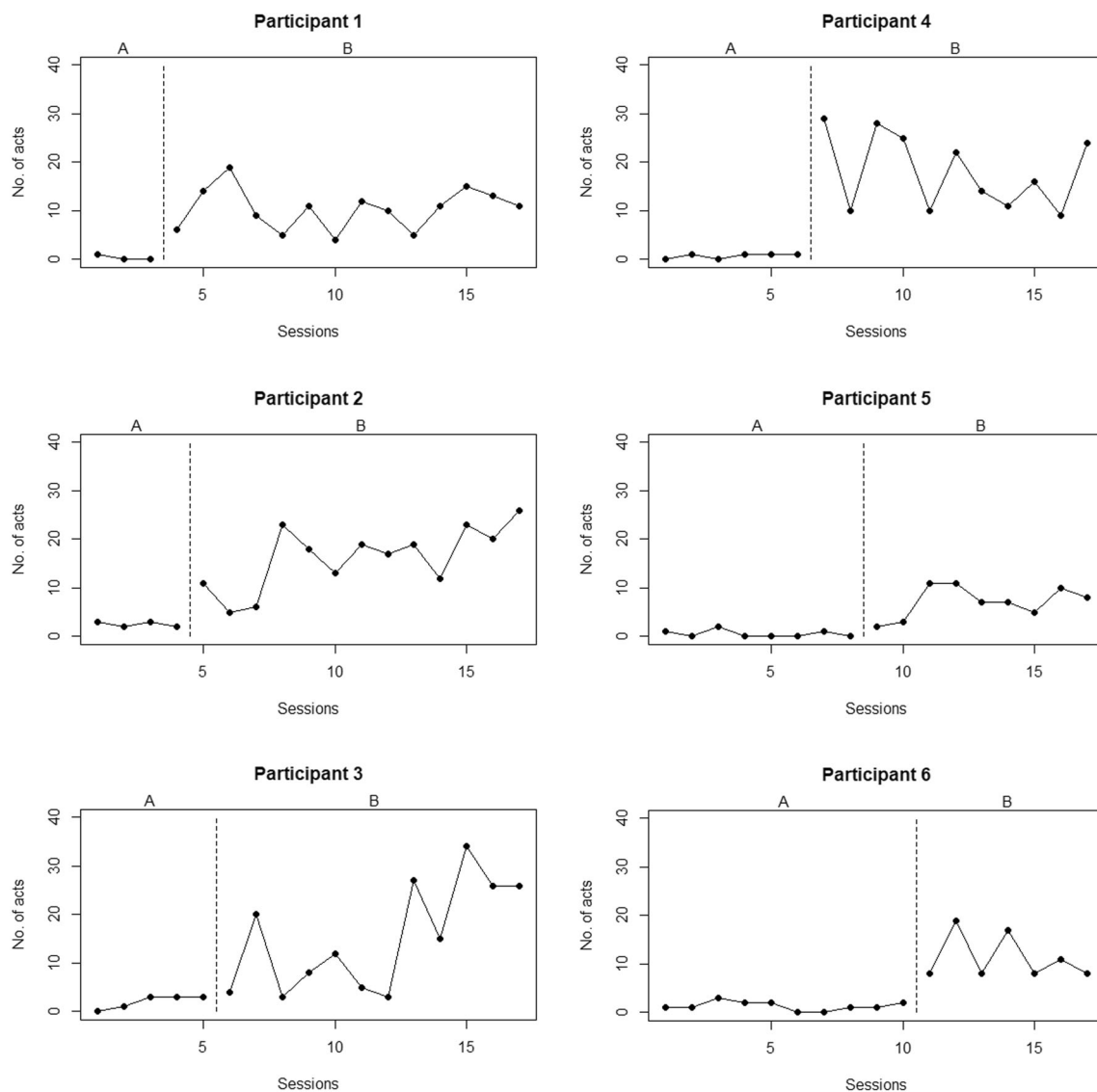


Fig. 1 Maximum numbers of child actions to maintain social interactions for 17 sessions in an MBD with six children in the study by Franco et al. (2013). The phases marked by “A” are the baseline phases and the phases marked by “B” are the treatment phases

restriction for computational simplicity and because this focus on complete data is consistent with the design logic of simultaneous replication and between-series comparisons in MBDs (Baer et al., 1968; Kazdin, 2011; Koehler & Levin, 2000). Analyses for unequal numbers of measurements for each case are possible in principle, but they would require more extensive coding and additional assumptions about the missing data. We will return to this limitation of equal numbers of repeated measurements for each case in the Discussion section.

Figure 1 displays the results for the first 17 sessions of the six children, with A denoting the baseline phase, in which no social interaction training was present, and B denoting the treatment phase, in which so-called prelinguistic milieu teaching techniques were applied (for details on these techniques, see Franco et al., 2013). Note the time-staggered way in which the treatment phase is started for the first up to the sixth child:

at Session 4, Session 5, Session 6, Session 7, Session 9, and Session 11, respectively.

With respect to the average treatment effect parameter (θ_{10}), the null hypothesis of the two-level model states that the value of this parameter would be 0. With the R script “example-data.R” provided in the Appendix, we can verify the observed value for this parameter for the two-level model described in Eqs. 1 and 2 using REML: $\theta_{10} = 11.9497$. The R script uses the `pbrktest` (Halekoh & Højsgaard, 2014) and `lmerTest` (Kuznetsova, Brockhoff, & Christensen, 2017) packages to compute the Kenward–Roger-adjusted F test, and it then gives the following output for the Franco et al. (2013) data:

This analysis-of-variance table shows that there is a statistically significant treatment effect if a 5% significance level is used, $F(1, 8.0906) = 37.49, p < .001$.

Note that although the Kenward–Roger approximation provides a more accurate (and conservative) hypothesis test for the average treatment effect across participants, in comparison to an unadjusted F test, it is still an F test, and as such the validity of its conclusions is dependent on the specific assumptions that are made for the application of F tests. These assumptions include random sampling, normally distributed errors at each level of the MLM, and equality of variances at each level of the MLM. However, research has shown that these assumptions are often not plausible in many domains of the social sciences, and especially not in single-case research (Micceri, 1989; Ruscio & Roche, 2012; Shadish & Sullivan, 2011; Solomon, 2014). An alternative way of making statistical inferences about the average treatment effect size in MBDs, without requiring any distributional assumptions or an assumption of random sampling, is by analyzing the data with an RT.

Analyzing multiple-baseline designs with randomization tests

RTs are nonparametric hypothesis tests that have been proposed for the analysis of MBDs (Bulté & Onghena, 2009; Koehler & Levin, 1998; Marascuilo & Busk, 1988; Wampold & Worsham, 1986). An RT can be used for statistical inference within a random-assignment model rather than a random-sampling model. In a random-assignment model, the statistical significance of a treatment effect can be determined by repartitioning the data a large number of times according to the randomization schedule of the randomized design and by calculating a test statistic S for each repartitioning of the data (Edgington & Onghena, 2007). This process yields a reference distribution that can be used for calculating nonparametric p -values (for small numerical examples and the underlying rationale, see, e.g., Heyvaert & Onghena, 2014; Levin et al., 2018; Onghena, 2018; Onghena, Tanious, De, & Michiels, 2019). By contrast, statistical tests within a random-sampling model (e.g., an F or t test) assume that the observed data were randomly sampled from a theoretical distribution (e.g., a normal distribution) in order to make valid statistical inferences about a treatment effect. In contrast, the RT does not make specific distributional assumptions or an assumption of random sampling, but obtains its validity from the randomization schedule that was actually used when designing the study. However, this also means that the use of RTs is only valid for single-case experimental designs (SCEDs) that incorporate some type of experimental random assignment. Note that the RT can be used with various type of test statistics, depending on the research question (Ferron & Sentovich, 2002; Onghena & Edgington, 2005).

Using the same empirical data as for the MLM example, we will now illustrate how MBDs can be randomized and analyzed with RTs. In that example, the start points of the

treatment for each child could have been randomly determined, taking into account that, in accordance with the What Works Clearinghouse Single-Case Design Standards (Kratochwill et al., 2010), each A phase and each B phase should contain at least three measurement occasions. Given these restrictions, the actual randomization can be performed by sampling the set of available start points (determined by the minimum number of measurement occasions per phase; in this case, Measurement Occasions 4–15) without replacement, and arranging them from small to large. This method has been called the “restricted Marascuilo–Busk procedure” by Levin et al. (2018), and this procedure performed well in their study in terms of Type I error rate control and power to detect immediate abrupt intervention effects. If s denotes the number of start points and n denotes the number of participants, then the number of possible randomizations equals $s!/(s - n)!$, given that $s \geq n$. In the example, this means that there are $12!/6! = 665,280$ possible randomizations.

Suppose that the randomly generated start points are 4, 5, 6, 7, 9, and 11 for Participants 1–6, respectively. Rather than using a theoretical reference distribution (such as a t or F distribution) to calculate a p value for the average treatment effect across cases, the RT uses an empirical reference distribution that is derived from the observed data. This is accomplished first by defining a measure that is sensitive to the effect that is expected (i.e., before observing the data), and second by calculating this measure for a large number of different start point randomizations on the observed data. Hence, the chosen measure operates as a test statistic, and the reference distributions consist of all values of this measure that could have been obtained if other start points were chosen (given a true null hypothesis).

For the present example, the treatment effect for each case is defined as the absolute mean difference between the A phase observations and the B phase observations. The average treatment effect for the complete MBD dataset is defined as the mean of the six treatment effects. With the R script “example-data.R” provided in the Appendix, we can verify that the observed value of this test statistic is 11.8353 in the present example. Next, we construct the empirical reference distribution for this observed value by calculating the selected test statistic for 4,999 different randomizations of treatment start point (the observed value of the test statistic is included in the reference distribution, bringing the total number of values to 5,000).³ After the reference distribution is derived, the two-sided p value can be defined as the proportion of test statistics in the randomization distribution that are at least as extreme as the observed test statistic. Using the R script “example-

³ In principle, 665,280 values can be calculated, but in order to keep the example computationally feasible, only 5,000 values were sampled. Such a Monte Carlo RT gives p values that are very close to the RT for the complete reference distribution (Edgington & Onghena, 2007).

data.R,” the two-sided p value is .008, indicating a statistically significant average treatment effect across cases at the 5% significance level. Note that this statistical inference is valid without requiring any of the assumptions that were required in the MLM example. However, we should emphasize that the inference is only valid when the start points of the MBD are indeed randomized.

Nonparametric inference for fixed effects in multilevel models: The randomization test wrapper

The rationale of the RT can be applied to any test statistic—hence, to any outcome of a statistical analysis, such as an analysis using an MLM (Cassell, 2002; Heyvaert et al., 2017; Onghena et al., 2018). If we apply the rationale to the tests used in an MLM, the MLM is repeatedly fitted to the data for K randomizations, and each time the resulting average treatment effect parameter of the MLM model fit is saved as an element of the reference distribution of the RT. Thus, the RT is wrapped around the MLM estimation procedure, hence the term “MLM-RT wrapper.” The result is a reference distribution of K average treatment effect parameters that can be used to assess the statistical significance of the observed average treatment effect. As before, the average treatment effect estimated with a two-level model for the example data is 11.9497. With the R script “example-data.R” provided in the Appendix, we can verify that for the MLM-RT wrapper, the two-sided p value for $K = 5,000$ is .0128. This is substantially larger than the p value for the conventional MLM, but still well below .05.

To more fully evaluate the added benefits of an RT wrapper procedure, the Type I error rate and the power must be assessed. In the remaining section of this article, we will evaluate the Type I error rate and power of the proposed RT wrapper by means of a simulation study (see, e.g., Peres-Neto & Olden, 2001, for a recommendation in this direction). In addition, we will compare the performance of the RT wrapper (in terms of Type I error rate and power) with the performance of parametric statistical inference in MLMs using F tests. Note that in the simulation study we will refer to the randomization test wrapper as “MLM-RT.”

Type I error rate and power of an MLM-RT wrapper to analyze randomized MBDs: A Monte Carlo simulation study

Method

A Monte Carlo simulation study was performed in which we manipulated the following simulation factors:

- *Number of cases:* 3, 4, 5, or 6.
- *Number of measurement occasions per case:* 20, 30, 40, or 50.
- *Distributional characteristics:* Data were generated from a standard normal distribution with an independent error structure, as in Eqs. 1 and 2; from a uniform distribution; from a first-order autoregressive model (AR1) with a positive autocorrelation of .6 and normally distributed residuals; or from a bimodal distribution (consisting out of two normal distributions with means of -2 and 2 , respectively). All distributions were set to have a mean of zero and a variance of 1, except for the AR1 model, in which the variance was slightly higher. More specifically, the variance of an AR1 model is $\frac{\sigma_e^2}{1-AR^2}$, where e is sampled from a standard normal distribution ($\sigma_e^2 = 1$). Given an AR value of .6, the variance of the AR1 model is 1.5625.
- *Size of the treatment effect:* 0, 0.5, 1, 1.5, 2, 2.5, or 3.
- *Between-case variance:* 1, 2, or 4.
- *Employed meta-analytic technique:* MLM or MLM-RT.

The significance level of all tests was set at 5%.

The number of cases in this simulation study was based on a survey by Shadish and Sullivan (2011), who found that for a large body of published single-case studies, the average number of cases in MBDs was 3.64. We chose three cases as a lower limit, because we obtained convergence problems for the MLM with the Kenward–Roger approximation for MBD datasets of fewer than three cases. The upper limit was chosen to be substantially larger than the lower limit in order to investigate the effect of the number of cases on the power, but still small enough to be relevant for single-case researchers, given the empirical average of 3.64 cases.

The number of measurement occasions for each case was also based on the results of Shadish and Sullivan (2011), who found that more than 90% of the surveyed studies contained 49 or fewer measurement occasions per case, with the median value being 20. On the basis of this observation, our selection of measurement occasions ranged from 20 to 50, with increments of ten measurement occasions. Note that the number of measurement occasions was kept constant across all cases within a single simulation condition.

Variability between the data of the individual cases within an MBD dataset was introduced by adding residual errors from a normal distribution with a mean of zero and a variance of 1, 2, or 4 to the null data of each case. In addition, the variance of the treatment effects was manipulated by generating treatment effects for every treatment measurement occasion in the MBD from a normal distribution with a mean equal to the average treatment effect but with differing variances (1, 2, or 4). It is worth mentioning that the random errors added to the data of the individual cases were uncorrelated with the random errors added to the treatment effects. Note also that

this way of generating the data yields datasets that contain more between-case variability than within-case variability. This was a deliberate choice, because research has shown that empirical datasets often contain considerably more between-case than within-case variability (Moeyaert, Ugille, Ferron, Beretvas, & Van den Noortgate, 2014).

With respect to the distributional characteristics of the generated data, a normal distribution was chosen in order to compare the performance of the MLM and MLM-RT for data in which the normality assumption of the MLM was not violated. The data from an AR1 model were included because research has shown that single-case data often contain positive autocorrelation (Shadish & Sullivan, 2011; Solomon, 2014). To keep the simulation study computationally manageable, we included only one value for the AR parameter. The autocorrelation of .6 was based on Shadish and Sullivan, who investigated the level of autocorrelation in published SCED data from various single-case designs. Their results showed that the level of autocorrelation ranged from near zero to .752, depending on the design. On the basis of Shadish and Sullivan's results, we chose .6 as a "bad-case scenario" value.

The uniform distribution and the bimodal distribution were included to account for two situations in which the normality assumption of the MLM is plainly violated. The uniform distribution was chosen because it is one of the simplest and most basic statistical distributions (Johnson, Kotz, & Balakrishnan, 1995) and because of its prominence as a benchmark in other simulation studies (Keller, 2012; Michiels, Heyvaert, & Onghena, 2018). In terms of observed data, it means simulating a condition in which there are no outliers, there is no distinct mode, and all scores within intervals of the same length are equally probable. The bimodal distribution was chosen because bimodality is common in behavioral research (Micceri, 1989) and because previous simulation studies had shown that standard techniques might lack robustness to deviations from unimodality (Keller, 2012; Poncet, Courvoisier, Combescure, & Perneger, 2016). In terms of observed data, it means simulating a condition in which two overlapping clusters of data unexpectedly appear, while keeping the number and size of the outliers comparable to the standard normal distribution. The bimodal distribution was generated by sampling data points from two normal distributions, with means of -2 and 2 , with equal probability. This level of mode separation was chosen in order to keep the modes clearly separated at all levels of between-case variability of the simulated data.

With regard to the size of the treatment effect, a wide range of values was chosen, with the upper limit being determined by a few single-case meta-analyses showing very high average treatment effect sizes of 3 or more (Fabiano et al., 2009; Heyvaert, Maes, Van den Noortgate, Kuppens, & Onghena, 2012; Heyvaert, Saenen, Campbell, Maes, & Onghena, 2014). This resulted in a selection of seven effect sizes, ranging from 0 to 3 in 0.5 increments.

Crossing all levels of all simulation factors resulted in 2,688 simulation conditions. For each condition, p values were calculated for 1,000 generated datasets. The power was defined as the proportion of p values that were equal to or smaller than .05 across all replications. For the MLM-RT technique, we used 1,000 random assignments to calculate a nonparametric p value for each replication. This implied 1,000,000 calculations for a single condition of MLM-RT.

Results

The results of the simulation study are summarized in Fig. 2. Note that in this figure, the y -axis label "Proportion of Rejections" is to be interpreted as the estimated power of the RT when the treatment effect was nonzero and as the Type I error rate when the treatment effect was zero.

Figure 2 shows a substantial difference in the average Type I error rates between MLM (left-hand panels) and MLM-RT (right-hand panels): Whereas the average Type I error rates of MLM are not controlled at all and can amount to 25% or more, MLM-RT has an average Type I error rate controlled at 5% or lower, as it should for a valid test at the 5% significance level. Closer inspection of Fig. 2a indicates that the main problem with the Type I error rate control for MLM occurs for the bimodal data, with an average Type I error rate above 75%.

With respect to estimated statistical power, all panels in Fig. 2 show the typical S-shaped power curve, in which the proportion of rejected null hypotheses increases monotonically when the treatment effect size increases. Figure 2a shows that, disregarding the bimodal data because of their grossly inflated Type I error rate, uniform data yield the highest power for both MLM and MLM-RT, followed by normal data, and then by data from an AR1 model.

Figure 2b shows that the power of both MLM and MLM-RT increases as the number of cases increases, but that this effect of number of cases is larger for MLM than for MLM-RT. Furthermore, even with the inflated Type I error rate, the power of the MLM is substantially lower than the power of the MLM-RT for three or four cases. For five and six cases with the uniform, normal, and autocorrelated datasets, this large power advantage for MLM-RT over MLM disappears.

In Fig. 2c, we can see that there is a positive relation between the number of measurement occasions per case and the power for both MLM and MLM-RT, as was expected. Over all conditions, the effect of the number of measurement occasions was smaller for MLM than for MLM-RT. From comparing panels b and c, we can see that the power of MLM is mostly determined by the number of cases and to a much lesser extent by the number of measurement occasions per case, whereas both factors matter about equally for MLM-RT.

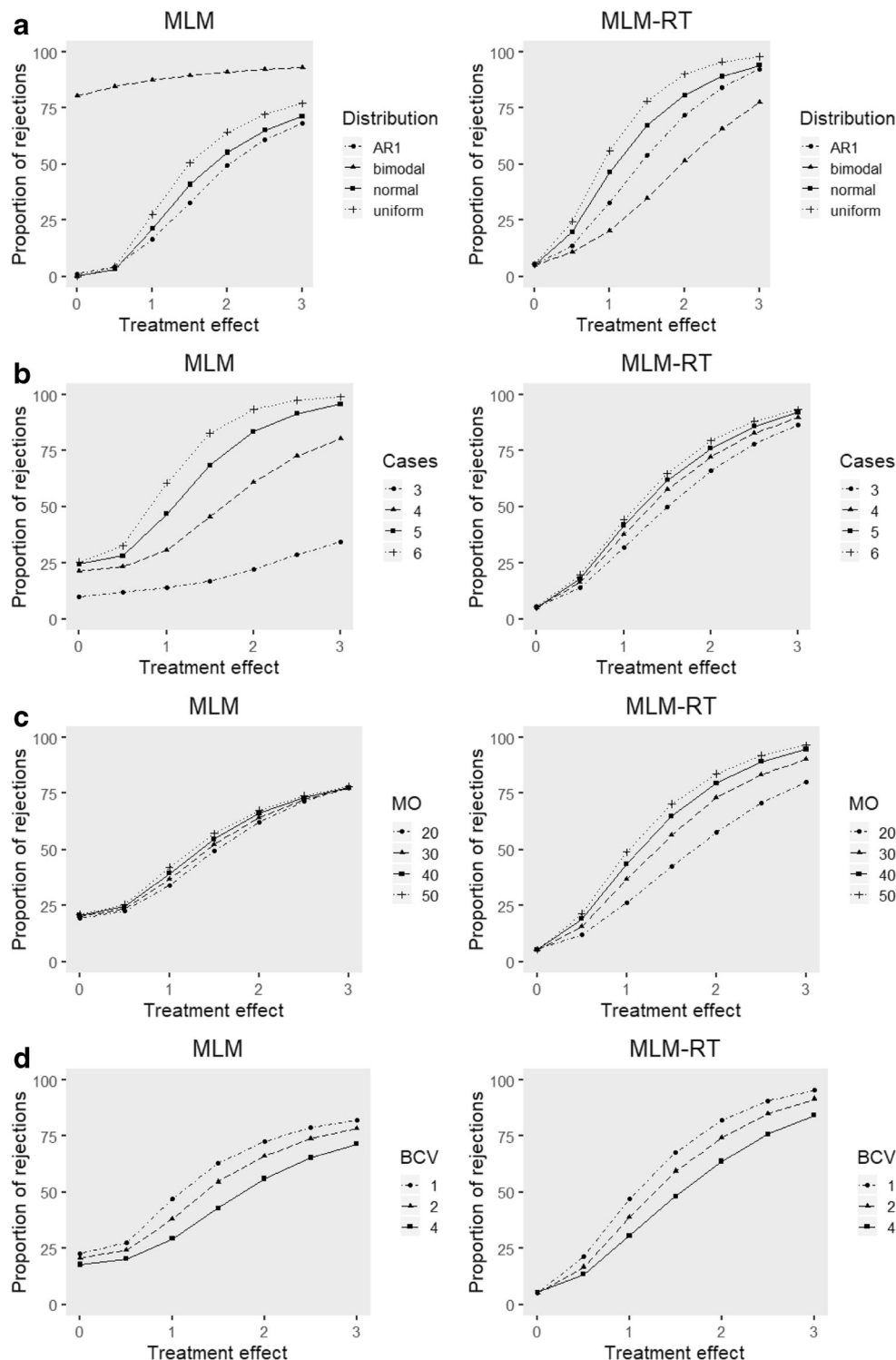


Fig. 2 Effects on the Type I error rates and power of MLM and MLM-RT of four different simulation factors (averaged over all other simulation factors). (a) The effects of data type. (b) The effects of the number of cases. (c) The effects of the number of measurement occasions (MO). (d)

Finally, Fig. 2d shows that the power of both MLM and MLM-RT is negatively affected by increasing between-case variance. The effects of between-case variance are similar for the two techniques.

The effects of the level of between-case variance (BCV). MLM = multi-level model, MLM-RT = multilevel model with randomization test, AR1 = first-order autoregressive model

In sum, the most important results of the simulation study are (1) that the Type I error rate of MLM-RT is controlled at the nominal significance level throughout the simulation study, while the Type I error rate of MLM is grossly inflated for

bimodal data, and (2) that MLM-RT has greater power than MLM for multiple-baseline datasets with three or four cases, and similar power for multiple-baseline datasets with five or six cases for the uniform, normal, and autocorrelated datasets.

Discussion

In this article, we have presented an MLM-RT wrapper in the context of analyzing data from single-case MBDs. First we gave an introduction to MLMs and demonstrated how they can be used to analyze data collected in MBDs. Second, we discussed and illustrated how RTs can be used as a nonparametric alternative for evaluating the average treatment effect across cases in an MBD. Third, we demonstrated how both approaches can be combined in order to make statistical inferences about the average treatment effect parameter of the MLM, without requiring distributional assumptions or an assumption of random sampling. Fourth, we evaluated the Type I error rate and power of both MLM and MLM-RT by means of a Monte Carlo simulation study.

The results of our Monte Carlo simulation study showed that the Type I error rate of MLM-RT was controlled at the nominal significance level for all conditions and that the Type I error rate of MLM using the Kenward–Roger adjustment for the degrees of freedom was controlled for the normal, uniform, and AR1 data, but not for bimodal data. These results for the MLM are in line with the simulation results obtained by Ferron et al. (2009; Ferron et al., 2010), but they also expand the previous results in a number of respects. Ferron et al. (2009; Ferron et al., 2010) simulated independent and autocorrelated data from a normal distribution and found that the use of the Kenward–Roger method led to accurate statistical inference about treatment effects, given the small numbers of participants that are common in MBDs. Complementary to this result, we found that for nonnormal distributions (e.g., bimodal data), grossly inflated Type I error rates for MLM are possible, even when using the Kenward–Roger adjustment for the degrees of freedom. This result showcases one of the downsides of the standard use of inferential statistical tests based on F distributions for single-case data. Furthermore, it illustrates that simulation results obtained under optimal conditions for a parametric test (simulating from a normal distribution) are not directly generalizable to more realistic conditions. The good news is that the use of an RT wrapper around MLM can remedy this problem with bimodality.

The power results of our Monte Carlo simulation study are also in line with the simulation results obtained by Levin et al. (2018), who examined the performance of randomization tests for MBD with different types of randomization schedules. They found Type I error rate control of the RT approach for normal and autocorrelated data and acceptable power values for small numbers of participants (smaller than six), just as in our simulation study. The result in our study that MBD data generated

from an AR1 model yield the smallest power for the MLM is also consistent with the MLM power literature (Ferron & Ware, 1995; Shadish, Kyse, & Rindskopf, 2013). Interestingly, we found that the power for both MLM and MLM-RT was largest for uniform MBD data. Although the simulation study by Michiels et al. (2018) had already shown that this is the case for RT-based analyses, one would not expect that this would be the case for MLM, given the model's assumption about normality of the residual errors. Further research will be needed to provide more insight into this effect.

In contrast to the previous simulation studies by Ferron et al. (2009; Ferron et al., 2010) and Levin et al. (2018), we included both MLM and an MLM-RT wrapper for our analysis of MBD data. Our results indicated that Type I error rate control is guaranteed better by the MLM-RT wrapper and that the power of the latter analysis is substantially greater than the power of regular MLM for MBD datasets with three and four cases. This is an important result, because a survey by Shadish and Sullivan (2011) showed that MBDs from published research use only an average of 3.64 cases. As such, we recommend that single-case researchers use MLM-RT whenever analyzing MBDs with fewer than five cases. Furthermore, we propose that the MLM-RT wrapper be used every time the parametric assumptions of MLM are considered implausible, regardless of the number of cases in the dataset.

Limitations and future research

We now discuss a few limitations of the present simulation study and propose future research avenues that can address these limitations. First, in the present simulation study the power of MLM and MLM-RT was only compared for the average treatment effect across all participants. MLMs also output various other parameters (e.g., various variance components or individual treatment effects) that can provide useful information regarding treatment effectiveness. Future research could focus on evaluating the power of MLM and the MLM-RT wrapper for various alternative MLM parameters. However, we should remark that not all MLM parameters are appropriate for use in MLM-RT. More specifically, because MLM-RT is based on the random assignment of experimental conditions in the SCED, only MLM parameters that pertain to the difference between the baseline phase and the treatment phase can be evaluated. For example, it is not possible to use MLM-RT for nonparametric inference with respect to between-case variance, because the experimental manipulation does not have an effect on this parameter. That being said, it is possible to use MLM-RT for evaluating differences in slope or differences in nonlinear effects between the baseline phase and the treatment phase by using more complex MLMs. An interesting avenue for future research could also be the modeling of delayed abrupt or immediate gradual intervention effects, as was examined by Levin,

Ferron, and Gafurov (2017) for RTs, and comparing the MLM and MLM-RT. In this context, it will also be relevant to investigate the performance of the MLM between-series estimator proposed by Ferron, Moeyaert, Van den Noortgate, and Beretvas (2014) and its robustness to bimodal distributions.

A second limitation is that the present simulation study only considered MBDs. Shadish, Kyse, and Rindskopf (2013) noted that MLMs can also be used to analyze data from other designs, such as replicated ABAB designs, alternating-treatment designs, and changing-criterion designs. An advantage of MLM-RT is that it can be used for any type of randomized SCED, as long as the randomization in the RT procedure mimics the type of randomization that was used in the (replicated) single-case experiment that is being analyzed. In this sense, future simulation studies could compare the power of MLM and MLM-RT for other designs.

A third limitation is that we only considered a specific instance of a two-level model (which is described in the introduction section), whereas there are many different ways in which the fixed and random effects in a two-level model can be specified. For this reason, future research should consider other instances of two-level models when comparing the power of MLM and MLM-RT. Furthermore, another avenue for further research is the application of MLM-RT to three-level models for the meta-analysis of multiple MBDs (e.g., Moeyaert et al., 2013).

A fourth limitation of this study is that we only simulated data from continuous and symmetrical distributions. Future research might focus on testing the generalizability of our results to discrete and/or skewed data. This should be particularly interesting and worthwhile, because Shadish and Sullivan (2011) found that more than 90% of the single-case designs in their review of the literature used some type of count variable as the outcome measure.

A fifth and final limitation is that we simulated complete datasets, so our findings are limited to sets without missing data and with equal case lengths. Future simulation studies will be needed to examine the performance of MLM-RT for datasets with missingness and unequal phase lengths.

This last limitation also holds for the R script “example-data.R,” provided in the Appendix, and for the illustrative analyses of the Franco et al. (2013) data presented in this article. If we want to extend MLM-RT to handle missing or incomplete data, two conditions would have to be met: (1) The multilevel model estimation procedure should be able to handle the missingness or incompleteness, and (2) the randomization procedure should avoid complete between-series exchangeability. With respect to the first condition, there are two main options: full maximum likelihood estimation (Fitzmaurice, Davidian, Verbeke, & Molenberghs, 2009; Snijders & Bosker, 2012) and multiple imputation (Peng & Chen, 2018; Sinharay, Stern, & Russell, 2001). However, for the inferences to remain valid, the data would have to be

missing at random and the (imputation) model should be correctly specified. For the second condition, Levin et al. (2018) recommended the restricted Marascuilo–Busk randomization procedure because, unlike other common randomization procedures for MBDs, it is based solely on within-series randomization.

In any case, if there are missing data or incomplete series, we should acknowledge that no statistical procedure exists that would magically make missing data reappear or make incomplete series complete again. Data analysis in practical research settings involving missingness or incompleteness should proceed cautiously and should combine information from multiple sources using multiple techniques. A sensible strategy might be to deal with the missing or incomplete data in different ways and to look for convergence or divergence in the resulting analyses. For example, in the Franco et al. (2013) data, besides truncation to the smallest series length, a simple alternative analysis could be based on the “last observation carried forward” procedure (White, Horton, Carpenter, & Pocock, 2011). Other simple procedures for dealing with missing data include mean imputation, linear inter- or extrapolation, or worst-case scenario imputation, although each of these procedures used separately is considered suboptimal (Schafer & Graham, 2002). Future research on missingness and unequal phase lengths in SCEDs should focus not only on the performance of the statistical techniques, but also on practical recommendations and interpretational caveats in empirical research with missing or incomplete data.

In the present article, we presented the MLM-RT only in the context of single-case research. In this respect, an interesting avenue for further research would be to apply MLM-RT to domains outside of single-case research (e.g., group-comparison research) in which different types of randomized experimental designs are used. Finally, we would emphasize that, although the results from the present simulation study indicate the merits of MLM-RT as compared to MLM in specific data-analytical situations, more research will be needed to validate the performance of MLM-RT with empirical data.

Software availability

The illustrations in this article can be reproduced using the R script “example-data.R” provided in the Appendix. The Appendix also contains R code for performing the simulation study. The randomization test wrapper proposed in this article is freely available via the following webpage: <https://ppw.kuleuven.be/mesrg/software-and-apps/mlm-rt>. Here, interested readers can download a .zip file that contains the relevant R code, along with a set of instructions on how to use the software and an example data file.

Author note This research was funded by the Research Foundation–Flanders (FWO), Belgium (grant ID: G.0593.14). The simulation study reported in this research was performed using the infrastructure of the VSC–Flemish Supercomputer Center, funded by the Hercules Foundation and the Flemish Government, Department EWI. The authors assure that all research presented in this article is fully original and has not been presented or made available elsewhere in any form.

References

- Alnahdi, G. H. (2015). Single-subject design in special education: Advantages and limitations. *Journal of Research in Special Educational Needs*, 15, 257–265. <https://doi.org/10.1111/1471-3802.12039>
- Baer, D. M., Wolf, M. M., & Risley, T. R. (1968). Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis*, 1, 91–97. <https://doi.org/10.1901/jaba.1968.1-91>
- Baldwin, S. A., & Fellingham, G. W. (2013). Bayesian methods for the analysis of small sample multilevel data with a complex variance structure. *Psychological Methods*, 18, 151–164. <https://doi.org/10.1037/a0030642>
- Barlow, D. H., Nock, M. K., & Hersen, M. (2009). Single case experimental designs: Strategies for studying behavior change (3rd). Boston: Pearson.
- Browne, W. J., Draper, D., Goldstein, H., & Rasbash, J. (2002). Bayesian and likelihood methods for fitting multilevel models with complex level-1 variation. *Computational Statistics and Data Analysis*, 39, 203–225. [https://doi.org/10.1016/S0167-9473\(01\)00058-5](https://doi.org/10.1016/S0167-9473(01)00058-5)
- Bulté, I., & Onghena, P. (2009). Randomization tests for multiple-baseline designs: An extension of the SCRT-R package. *Behavior Research Methods*, 41, 477–485. <https://doi.org/10.3758/BRM.41.2.477>
- Burrick, R. K., & Graybill, F. A. (1992). Confidence intervals on variance components. New York: Marcel Dekker.
- Cassell, D. L. (2002). A randomization-test wrapper for SAS® PROCs. *SAS User's Group International Proceedings*, 27, 251. Retrieved from <http://www.lexjansen.com/wuss/2002/WUSS02023.pdf>
- Clarke, P., & Wheaton, B. (2007). Addressing data sparseness in contextual population research using cluster analysis to create synthetic neighborhoods. *Sociological Methods & Research*, 35, 311–351. <https://doi.org/10.1177/0049124106292362>
- Edgington, E. S. (1969). Statistical inference: The distribution-free approach. New York: McGraw-Hill.
- Edgington, E. S. (1996). Randomized single-subject experimental designs. *Behaviour Research and Therapy*, 34, 567–574. [https://doi.org/10.1016/0005-7967\(96\)00012-5](https://doi.org/10.1016/0005-7967(96)00012-5)
- Edgington, E. S., & Onghena, P. (2007). Randomization tests (4th). Boca Raton: Chapman & Hall/CRC.
- Fabiano, G. A., Pelham, W. E., Coles, E. K., Gnagy, E. M., Chronis-Tuscano, A., & O'Connor, B. C. (2009). A meta-analysis of behavioral treatments for attention-deficit/hyperactivity disorder. *Clinical Psychology Review*, 29, 129–140. <https://doi.org/10.1016/j.cpr.2008.11.001>
- Fai, A. H. T., & Cornelius, P. L. (1996). Approximate F -tests of multiple degree of freedom hypotheses in generalized least squares analyses of unbalanced split-plot experiments. *Journal of Statistical Computing and Simulation*, 54, 363–378. <https://doi.org/10.1080/00949659608811740>
- Fedorov, S. (2013). GetData graph digitizer. Retrieved from <http://getdata-graphdigitizer.com/>
- Ferron, J., & Ware, W. (1995). Analyzing single-case data: The power of randomization tests. *Journal of Experimental Education*, 63, 167–178. <https://doi.org/10.1080/00220973.1995.9943820>
- 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, 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, 930–943. <https://doi.org/10.3758/BRM.42.4.930>
- Ferron, J. M., Moeyaert, M., Van den Noortgate, W., & Beretvas, S. N. (2014). Estimating causal effects from multiple-baseline studies: Implications for design and analysis. *Psychological Methods*, 19, 493–510. <https://doi.org/10.1037/a0037038>
- Ferron, J. M., & Sentovich, C. (2002). Statistical power of randomization tests used with multiple-baseline designs. *Journal of Experimental Education*, 70, 165–178. <https://doi.org/10.1080/00220970209599504>
- Fitzmaurice, G. M., Davidian, M., Verbeke, G., & Molenberghs, G. (2009). Longitudinal data analysis. Boca Raton: Chapman & Hall/CRC.
- Franco, J. H., Davis, B. L., & Davis, J. L. (2013). Increasing social interaction using prelinguistic milieu teaching with nonverbal school-age children with autism. *American Journal of Speech-Language Pathology*, 22, 489–502. [https://doi.org/10.1044/1058-0360\(2012/10-0103\)](https://doi.org/10.1044/1058-0360(2012/10-0103))
- Gabler, N. B., Duan, N., Vohra, S., & Kravitz, R. L. (2011). N -of-1 trials in the medical literature: A systematic review. *Medical Care*, 49, 761–768. <https://doi.org/10.1097/MLR.0b013e318215d90d>
- Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1, 515–533.
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2013). Bayesian data analysis (3rd). Boca Raton: Chapman and Hall/CRC.
- Halekoh, U., & Højsgaard, S. (2014). A Kenward–Roger approximation and parametric bootstrap methods for tests in linear mixed models: The R package pbkrtest. *Journal of Statistical Software*, 59, 1–32.
- Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72, 320–340. <https://doi.org/10.1080/01621459.1977.10480998>
- Heyvaert, M., Maes, B., Van den Noortgate, W., Kuppens, S., & Onghena, P. (2012). A multilevel meta-analysis of single-case and small- n research on interventions for reducing challenging behavior in persons with intellectual disabilities. *Research in Developmental Disabilities*, 33, 766–780. <https://doi.org/10.1016/j.ridd.2011.10.010>
- Heyvaert, M., Moeyaert, M., Verkempynck, P., Van den Noortgate, W., Vervloet, M., Ugille, M., & Onghena, P. (2017). Testing the intervention effect in single-case experiments: A Monte Carlo simulation study. *Journal of Experimental Education*, 85, 175–196. <https://doi.org/10.1080/00220973.2015.1123667>
- Heyvaert, M., & Onghena, P. (2014). Randomization tests for single-case experiments: State of the art, state of the science, and state of the application. *Journal of Contextual Behavioral Science*, 3, 51–64. <https://doi.org/10.1016/j.jcbs.2013.10.002>
- Heyvaert, M., Saenen, L., Campbell, J. M., Maes, B., & Onghena, P. (2014). Efficacy of behavioral interventions for reducing problem behavior in persons with autism: An updated quantitative synthesis of single-subject research. *Research in Developmental Disabilities*, 35, 2463–2476. <https://doi.org/10.1016/j.ridd.2014.06.017>
- Heyvaert, M., Wendt, O., Van den Noortgate, W., & Onghena, P. (2015). Randomization and data-analysis items in quality standards for

- single-case experimental studies. *Journal of Special Education*, 49, 146–156. <https://doi.org/10.1177/0022466914525239>
- Jenson, W. R., Clark, E., Kircher, J. C., & Kristjansson, S. D. (2007). Statistical reform: Evidence-based practice, meta-analyses, and single subject designs. *Psychology in the Schools*, 44, 483–493. <https://doi.org/10.1002/pits.20240>
- Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). Continuous univariate distributions, Vol. 2 (2nd). New York: Wiley.
- Kazdin, A. E. (2011). Single-case research designs: Methods for clinical and applied settings (2nd). New York: Oxford University Press.
- Keller, B. (2012). Detecting treatment effects with small samples: The power of some tests under the randomization model. *Psychometrika*, 2, 324–338. <https://doi.org/10.1007/s11336-012-9249-5>
- Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 53, 983–997. <https://doi.org/10.2307/2533558>
- Kenward, M. G., & Roger, J. H. (2009). An improved approximation to the precision of fixed effects from restricted maximum likelihood. *Computational Statistics and Data Analysis*, 53, 2583–2595. <https://doi.org/10.1016/j.csda.2008.12.013>
- Koehler, M. J., & Levin, J. R. (1998). Regulated randomization: A potentially sharper analytical tool for the multiple-baseline design. *Psychological Methods*, 3, 206–217. <https://doi.org/10.1037/1082-989X.3.2.206>
- Koehler, M. J., & Levin, J. R. (2000). RegRand: Statistical software for the multiple-baseline design. *Behavior Research Methods, Instruments, & Computers*, 32, 367–371. <https://doi.org/10.3758/BF03207807>
- Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). *Single-case designs technical documentation*. Retrieved from What Works Clearinghouse website: http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf
- Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods*, 15, 124–144. <https://doi.org/10.1037/a0017736>
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13), 1–26.
- Levin, J. R., Ferron, J. M., & Gafurov, B. S. (2017). Additional comparisons of randomization-test procedures for single-case multiple-baseline designs: Alternative effect types. *Journal of School Psychology*, 63, 13–34. <https://doi.org/10.1016/j.jsp.2017.02.003>
- Levin, J. R., Ferron, J. M., & Gafurov, B. S. (2018). Comparison of randomization-test procedures for single-case multiple-baseline designs. *Developmental Neurorehabilitation*, 21, 290–311. <https://doi.org/10.1080/17518423.2016.1197708>
- Levin, J. R., O'Donnell, A. M., & Kratochwill, T. R. (2003). Educational/psychological intervention research. In I. B. Weiner (Series Ed.), W. M. Reynolds & G. E. Miller (Vol. Eds.), *Handbook of psychology: Vol. 7. Educational psychology* (pp. 557–581). Hoboken, NJ: Wiley.
- Maas, C. J. M., & Hox, J. J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica*, 58, 127–137. <https://doi.org/10.1046/j.0039-0402.2003.00252.x>
- Marascuilo, L. A., & Busk, P. L. (1988). Combining statistics for multiple-baseline AB and replicated ABAB designs across subjects. *Behavioral Assessment*, 10, 1–28.
- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105, 156–166. <https://doi.org/10.1037/0033-2909.105.1.156>
- Michiels, B., Heyvaert, M., & Onghena, P. (2018). The conditional power of randomization tests for single-case effect sizes in designs with randomized treatment order: A Monte Carlo simulation study. *Behavior Research Methods*, 50, 557–575. <https://doi.org/10.3758/s13428-017-0885-7>
- 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, 760–778. <https://doi.org/10.1037/met0000136>
- Moeyaert, M., Ugille, M., Ferron, J. M., Beretvas, S. N., & Van den Noortgate, W. (2013). The three-level synthesis of standardized single-subject experimental data: A Monte Carlo simulation study. *Multivariate Behavioral Research*, 48, 719–748. <https://doi.org/10.1080/00273171.2013.816621>
- Moeyaert, M., Ugille, M., Ferron, J. M., Beretvas, S. N., & Van den Noortgate, W. (2014). Three-level analysis of single-case experimental data: Empirical validation. *Journal of Experimental Education*, 82, 1–21. <https://doi.org/10.1080/00220973.2012.745470>
- Onghena, P. (2005). Single-case designs. In B. Everitt & D. Howell (Eds.), *Encyclopedia of statistics in behavioral science*, vol. 4 (pp. 1850–1854). Chichester, UK: Wiley.
- Onghena, P. (2018). Randomization and the randomization test: Two sides of the same coin. In V. Berger (Ed.), *Randomization, masking, and allocation concealment* (pp. 185–207). Boca Raton: Chapman & Hall/CRC Press.
- Onghena, P., & Edgington, E. S. (2005). Customization of pain treatments: Single-case design and analysis. *Clinical Journal of Pain*, 21, 56–68.
- Onghena, P., Michiels, B., Jamshidi, L., Moeyaert, M., & Van den Noortgate, W. (2018). One by one: Accumulating evidence by using meta-analytical procedures for single-case experiments. *Brain Impairment*, 19, 33–58. <https://doi.org/10.1017/BrImp.2017.25>
- Onghena, P., Tanius, R., De, T. K., & Michiels, B. (2019). Randomization tests for changing criterion designs. *Behaviour Research and Therapy*, 117, 18–27. <https://doi.org/10.1016/j.brat.2019.01.005>
- Peng, C.-Y. J., & Chen, L.-T. (2018). Handling missing data in single-case studies. *Journal of Modern Applied Statistical Methods*, 17, eP2488. <https://doi.org/10.22237/jmasm/1525133280>
- Peres-Neto, P. R., & Olden, J. D. (2001). Assessing the robustness of randomization tests: Examples from behavioural studies. *Animal Behaviour*, 61, 79–86. <https://doi.org/10.1006/anbe.2000.1576>
- Poncet, A., Courvoisier, D. S., Combescur, C., & Perneger, T. V. (2016). Normality and sample size do not matter for the selection of an appropriate statistical test for two-group comparisons. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 12, 61–71. <https://doi.org/10.1027/1614-2241/a000110>
- Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods (Vol. 2). London: Sage.
- Ruscio, J., & Roche, B. (2012). Variance heterogeneity in published psychological research: A review and a new index. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 8, 1–11. <https://doi.org/10.1027/1614-2241/a000034>
- Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika*, 6, 309–316. <https://doi.org/10.1007/BF02288586>
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7, 147–177. <https://doi.org/10.1037/1082-989X.7.2.147>
- Shadish, W. R., Hedges, L. V., Pustejovsky, J., Rindskopf, D. M., Boyajian, J. G., & Sullivan, K. J. (2014). Analyzing single-case designs: *d*, *G*, multilevel models, Bayesian estimators, generalized additive models, and the hopes and fears of researchers about analysis. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and data-analysis advances* (pp. 247–281). Washington, DC: American Psychological Association.
- Shadish, W. R., Kyse, E. N., & Rindskopf, D. M. (2013). Analyzing data from single-case designs using multilevel models: New applications

- and some agenda items for future research. *Psychological Methods*, 18, 385–405. <https://doi.org/10.1037/a0032964>
- Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods*, 43, 971–980. <https://doi.org/10.3758/s13428-011-0111-y>
- Sinharay, S., Stern, H. S., & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods*, 6, 317–329. <https://doi.org/10.1037/1082-989X.6.4.317>
- Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd). London: Sage.
- Solomon, B. G. (2014). Violations of assumptions in school-based single-case data: Implications for the selection and interpretation of effect sizes. *Behavior Modification*, 38, 477–496. <https://doi.org/10.1177/0145445513510931>
- Spiegelhalter, D. J., Abrams, K. R., & Myles, J. P. (Eds.). (2004). *Bayesian approaches to clinical trials and health-care evaluation*, Chichester: Wiley.
- Swaminathan, H., & Rogers, H. J. (2007). Statistical reform in school psychology research: A synthesis. *Psychology in the Schools*, 44, 543–549. <https://doi.org/10.1002/pits.20246>
- Tyrrell, P. N., Corey, P. N., Feldman, B. M., & Silverman, E. D. (2013). Increased statistical power with combined independent randomization tests used with multiple-baseline design. *Journal of Clinical Epidemiology*, 66, 691–694. <https://doi.org/10.1016/j.jclinepi.2012.11.006>
- Van den Noortgate, W., & Onghena, P. (2003a). Combining single case experimental studies using hierarchical linear models. *School Psychology Quarterly*, 18, 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, 1–10. <https://doi.org/10.3758/BF03195492>
- Van den Noortgate, W., & Onghena, P. (2007). The aggregation of single-case results using hierarchical linear models. *Behavior Analyst Today*, 8, 196–209. <https://doi.org/10.1037/h0100613>
- Van den Noortgate, W., & Onghena, P. (2008). A multilevel meta-analysis of single-subject experimental design studies. *Evidence Based Communication Assessment and Intervention*, 2, 142–151. <https://doi.org/10.3758/s13428-012-0213-1>
- Wampold, B. E., & Worsham, N. L. (1986). Randomization tests for multiple-baseline designs. *Behavioral Assessment*, 8, 135–143.
- White, I. R., Horton, N. J., Carpenter, J., & Pocock, S. J. (2011). Strategy for intention to treat analysis in randomized trials with missing outcome data. *BMJ*, 342, d40. <https://doi.org/10.1136/bmj.d40>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.