Testing the Intervention Effect in Single-Case Experiments:

A Monte Carlo Simulation Study

Mieke Heyvaert[1, 2], Mariola Moeyaert[1, 2], Paul Verkempynck, Wim Van den Noortgate[1], Marlies Vervloet[1], Maaike Ugille[1], & Patrick Onghena[1]

[1] Faculty of Psychology and Educational Sciences - KU Leuven

[2] Postdoctoral Fellow of the Research Foundation - Flanders (Belgium)

Correspondence concerning this article can be addressed to Dr. Mieke Heyvaert, Methodology of Educational Sciences Research Group, Tiensestraat 102 - Box 3762, B-3000 Leuven, Belgium. Phone +32 16 326265. E-mail Mieke.Heyvaert@ppw.kuleuven.be

Testing the Intervention Effect in Single-Case Experiments: A Monte Carlo Simulation Study

Abstract: This article reports on a Monte Carlo simulation study, evaluating two approaches for testing the intervention effect in replicated randomized AB designs: two-level hierarchical linear modeling (HLM) and using the additive method to combine randomization test *p* values (RTcombiP). Four factors were manipulated: the mean intervention effect, the number of cases included in a study, the number of measurement occasions for each case, and the between-case variance. Under the simulated conditions, Type I error rate was under control at the nominal 5% level for both HLM and RTcombiP. Furthermore, for both procedures, a larger number of combined cases resulted in higher statistical power, with many realistic conditions reaching statistical power of 80% or higher. Smaller values for the between-case variance resulted in higher power for HLM. A larger number of data points resulted in higher power for RTcombiP.

Keywords: Hierarchical linear models, Randomization tests, Single-case experimental design, Statistical power, Type I error rate

A considerable number of studies published in the domain of education rely on single-case experimental designs (SCEDs) (Shadish & Sullivan, 2011). SCEDs can be used to evaluate the effect of an intervention for a single entity by comparing the repeated measurements of a dependent variable under at least two manipulated conditions, typically a baseline and a treatment condition. The most basic SCEDs are AB designs: Repeated measurements of the dependent variable are first made under a control condition in the A phase or baseline phase, and then continued under an experimental condition in the B phase or intervention phase. Because all baseline measurements precede all treatment measurements, AB designs are suitable to study irreversible behavior or behavior that is undesirable to return to baseline conditions for ethical or practical reasons. For instance, when in the B phase an intervention was introduced that targeted reading skills, math skills, or social skills, it is unlikely that all treatment effects would disappear when the intervention is discontinued in the withdrawal phase in an ABA design (i.e., the second A phase). Furthermore, it can be considered unethical to withdraw a beneficial intervention and re-introduce an A phase.

A major disadvantage of AB designs is that they provide little control over internal validity threats, such as history and maturation. 'History' refers to the influence of *external* events (e.g., weather change, big news event, holidays) that occur during the course of an SCED that may influence the participant's behavior in such a way as to make it appear that there was a treatment effect, whereas 'maturation' deals with changes *within* the participant (e.g., physical maturation, tiredness, boredom, hunger) during the course of the SCED (Edgington, 1996).

A first possible answer to internal validity threats is to include randomization in SCEDs. In SCEDs where randomization is feasible and logical, random assignment of the measurement occasions to the experimental conditions can yield statistical control over (known and unknown) confounding variables and can facilitate causal inference (Heyvaert,

Wendt, Van den Noortgate, & Onghena, in press; Kratochwill & Levin, 2010; Onghena & Edgington, 2005). In AB designs, the start of the intervention (i.e., the moment of phase change) can be randomly determined. In order to prevent that the number of measurement occasions for a phase would be too small, it is usually recommended using a restricted randomized phase change, in which a minimum length for each phase is determined *a priori*. For instance, a randomized AB design with eight measurement occasions and with at least three measurement occasions in each phase implies three possible assignments: AAABBBBB, AAAABBBB, and AAAAABBB. Next, one of the possible assignments is selected randomly (cf. Bulté & Onghena, 2008).

A second possible answer to validity threats is to include replication in SCEDs. SCEDs can be replicated simultaneously or sequentially. The multiple baseline across participants design is an often used simultaneous replication design: Several AB designs are conducted at the same time over several participants, and the intervention is introduced at different moments in time for the included participants in order to control for historical confounding factors. An alternative to the basic simultaneous replication design is the *randomized* simultaneous replication design: For each participant the moment of phase change is randomly determined, while simultaneous phase change for the included participants is avoided (cf. Bulté & Onghena, 2009). Using simultaneous replication designs can be challenging. First, they can involve a high workload for the experimenter because all data have to be collected in the same period for all included participants. Second, the intervention is withheld temporarily from some participants in order to assure the staggered administration of the intervention, which might imply ethical (e.g., withholding effective treatment) and practical (e.g., boredom) problems. An alternative to simultaneous replication designs are sequential replication designs: The replications over the included participants are carried out consecutively. In *randomized* sequential replication designs, for each participant

the moment of phase change is randomly determined. For instance, in a randomized sequential AB replication design with four participants, for each participant the start of the B phase is randomly determined *a priori*, then the experiments are conducted consecutively for the four participants, and afterwards the collected data are analyzed over the four participants. In our simulation study we will focus on testing intervention effects in randomized sequential replication designs. Empirical examples of randomized sequential replication designs are for instance Holden et al. (2003), ter Kuile et al. (2009), O'Neill and Findlay (2014), Van de Vliet et al. (2003), and Vlaeyen, de Jong, Geilen, Heuts, and van Breukelen (2001).

Testing the intervention effect in randomized sequential replication designs

A first step in analyzing SCE data is visual analysis (Egel & Barthold, 2010; Kratochwill et al., 2010). Single-case researchers are advised to examine the following features and data patterns within and between the phases of their SCED: (1) level, (2) trend, (3) variability, (4) immediacy of the effect, (5) overlap, and (6) consistency of data patterns across similar phases (see e.g., Bulté & Onghena, 2012, and Kratochwill et al., 2010, for guidelines and available software).

Second, researchers can statistically test the intervention effect. This testing of the intervention effect will be the focus of the present paper. Various parametric and nonparametric approaches could be used for testing the intervention effect in randomized sequential replication designs. In our simulation study we will focus on two-level hierarchical linear modeling (HLM; Van den Noortgate & Onghena, 2003a, 2003b), a parametric approach, and using the additive method to combine randomization test *p* values (RTcombiP; Edgington & Onghena, 2007), a nonparametric approach.

The first approach we studied was the two-level HLM approach of Van den Noortgate and Onghena (2003a, 2003b). HLM is opportune for analyzing hierarchically structured data.

In randomized sequential replication designs, the measurement occasions are nested within the participants. This implies that the measurements for one participant are probably more alike than the measurements for another participant included in the replication design. HLM allows taking into account the dependencies that may result from the hierarchical clustering. Using two-level HLM for analyzing randomized sequential replication designs allows to estimate various parameters of interest: case-specific intercepts and treatment effects, the average baseline level over the included cases, the average treatment effect over the included cases, and estimates for within- and between-case variance in the baseline level and in the treatment effect. The HLM approach typically includes using the Wald test to test the null hypothesis that on average there is no statistically significant effect of the independent variable on the level of the dependent variable, and using the likelihood ratio test to test the variance at the between-case level. When the latter test indicates that there is significant between-case variance, the presence of moderators is likely. Predictor variables can be included in the analyses to test whether the treatment effect depends on characteristics of the cases included in the replication design. The HLM approach for analyzing randomized sequential replication designs is related to other regression approaches proposed for analyzing SCED data, for instance ordinary least squares regression analysis (Huitema & McKean, 1998), generalized least squares regression analysis (Maggin et al., 2011), interrupted time series analysis procedures such as ITSACORR (Crosbie, 1993, 1995), and piecewise regression analysis (Center, Skiba, & Casey, 1985-1986). Whereas these approaches however are developed for the analysis of data from one single case, the HLM approach allows combining the data from multiple cases, accounting for the dependencies that may result from the hierarchical clustering of the SCED data.

Our simulation study on testing the intervention effect in randomized sequential replication designs concerns two-level HLMs. However, it is also possible to add an

additional level and use three-level HLMs, for instance when conducting a meta-analysis of published SCED studies (see e.g., Moeyaert, Ugille, Ferron, Beretvas, & Van den Noortgate, 2014). For the present simulation study we used a two-level HLM, as described by Van den Noortgate and Onghena (2003a, 2003b), for testing the intervention effect. The restricted maximum likelihood estimation (REML) approach in SAS PROC MIXED was used to estimate the overall intervention effect (Littell, Milliken, Stroup, Wolfinger, & Schabenberger, 2006). We used the Satterthwaite approach to approximate the degrees of freedom and to derive the corresponding *p* value, because this approach showed to provide accurate confidence intervals for the estimates of the average treatment effect for the two-level analysis of SCED data (Ferron, Bell, Hess, Rendina-Gobioff, & Hibbard, 2009).

In addition to this specific two-level HLM approach, other HLM approaches and extensions have been developed and used for analyzing randomized sequential replication designs. For instance, two-level HLMs can be extended to account for trends (linear and non-linear; Shadish, Kyse, & Rindskopf, 2013; Van den Noortgate & Onghena, 2003b), autocorrelation (Van den Noorgate & Onghena, 2003a), unequal within-phase variances (Baek & Ferron, 2013), external events (Moeyaert, Ugille, Ferron, Beretvas, & Van den Noortgate, 2013b), and non-normal outcomes, such as counts (Shadish et al., 2013). Furthermore, as an alternative to using the REML approach for estimating the overall intervention effect (Littell et al., 2006), maximum likelihood (ML) estimations (e.g., Moeyaert, Ferron, Beretvas, & Van den Noortgate, 2014) and Bayesian methods (e.g., Shadish et al., 2013) have also been applied in the context of SCEDs. We acknowledge that the way we have operationalized the two-level HLM approach with a specific model and with a specific estimation method may impact the results and conclusions of our simulation study (cf. *Discussion*).

The second approach we studied was the RTcombiP approach of Edgington and Onghena (2007). This RTcombiP approach for testing the intervention effect in randomized sequential replication designs includes first conducting a randomization test (RT) for each participant, and afterwards using the additive method to combine the RT *p* values for all included participants. The RT is a statistical significance test based on the random assignment of the measurement occasions to the experimental conditions (Edgington & Onghena, 2007; Koehler & Levin, 1998; Levin & Wampold, 1999; Todman & Dugard, 2001; Wampold & Worsham, 1986). In order to guarantee the validity of RTs, it is generally advised to randomly assign the measurement occasions to the experimental conditions before the start of the SCED (see *Discussion* for exceptions and alternative procedures). Furthermore, it is generally advised to formulate the null hypothesis and the alternative hypothesis and select an appropriate test statistic *a priori* to conducting the experiment (see *Discussion* for alternative procedures). After collecting the empirical data for each participant, the RT involves calculating the test statistic for each possible assignment and looking where the observed test statistic (i.e., the value of the test statistic based on the collected data) falls within the distribution of all possible test statistic values (i.e., the randomization distribution). The *p* value of the RT is calculated as the proportion of possible test statistic values that is as extreme, or even more extreme, than the observed test statistic. A second step is to combine the RT *p* values for all participants included in the randomized sequential replication design. For our simulation study we will use the additive method for determining the significance of a sum of *p* values (see Edgington, 1972). In the RTcombiP approach, the overall null hypothesis is that there is no treatment effect for any of the cases included in the study. An advantage of the RTcombiP approach is that it does not make any assumptions about the distribution of the SCED data and does not demand that the SCED data are serially independent.

In addition to the RTcombiP approach of Edgington and Onghena (2007) we used for the present simulation study, other RT approaches have been developed. Whereas the RTcombiP approach for testing the intervention effect in randomized sequential replication designs includes first conducting an RT for each participant and afterwards using the additive method to combine the RT *p* values for all participants included in the SCED, the Marascuilo and Busk (1988) RT approach, for instance, works with the raw data instead of the *p* values. In the *Discussion* section, we will focus on how the selected RT approach may impact the results and conclusions of the simulation study.

In the following sections we will present an empirical example to illustrate how single-case researchers can use HLM and RTcombiP to test the intervention effect in randomized sequential replication designs. The empirical example will be helpful for understanding the design and the methods of our Monte Carlo simulation study. After presenting the empirical example, we will describe the aims, methods, and results of our simulation study.

Empirical illustration: Testing the intervention effect for one participant within a randomized AB design

Our first fictive example concerns a randomized AB design that is used to evaluate the effect of an intervention aimed at increasing the self-esteem of a student. The self-esteem is measured daily on a scale from 0 to 10. There are 30 measurement occasions. In accordance with the *Design and Evidence Standards* for SCEDs developed by the *What Works Clearinghouse* (Kratochwill et al., 2010), the experimenter wishes to have minimally three observations per phase. This implies that there are 25 possible assignments. Randomly, one out of the possible assignments is selected, and the actual experiment has to be conducted in accordance with this assignment. For our example, the random assignment resulted for this participant in 11 measurement occasions during the A phase and 19 measurement occasions

during the B phase (cf. first column of Table 1). The self-esteem scores for this participant ('Participant 1') are described in the second column of Table 1, and plotted in Figure 1.

Using ordinary least square regression for testing the intervention effect for one participant. A parametric approach that can be used to test the intervention effect for this participant is ordinary least square regression (OLS) analysis (the HLM approach we will focus on in our simulation study is an extension of the OLS approach). The equation that can be used to conduct this analysis is:

Eq. (1): $y_i = \beta_0 + \beta_1 Phase_i + e_i$

In this regression equation, the measurement occasions are depicted by the symbol *i*. The dummy coded variable $Phase_i$ indicates the experimental condition: If $Phase_i$ = 0 then measurement occasion *i* belongs to the baseline phase, and if $Phase_i$ = 1 then *i* belongs to the intervention phase. β_0 and β_1 indicate the intercept (i.e., baseline level) and treatment effect (difference between baseline level and level in the treatment phase), respectively. The model assumes that the within-case residuals (e_i) are normally distributed.

We used SAS® 9.3 Software (SAS Institute Inc., 2011-2015) to conduct the OLS analyses. Appendix A includes the SAS code for analyzing this dataset. PROC REG was used to estimate the baseline level (i.e., $\hat{\beta}_0$) and the treatment effect (i.e., $\hat{\beta}_1$) for this student. For both parameters, SAS conducted a Student *t* test of the null hypothesis that the true parameter is zero. The estimated baseline level for this student was 1.55 (*SE* = 0.15), $t(28) = 10.00$, $p < .0001$, indicating a low level of self-esteem in the baseline phase that is statistically significant at an alpha level of .05. The estimated treatment effect for this student was 4.03 (*SE* = 0.19), $t(28) = 20.77$, $p < .0001$, indicating an increase in self-esteem due to the treatment this is statistically significant at an alpha level of .05.

Using the RT for testing the intervention effect for one participant. A nonparametric approach that can be used to test the intervention effect for this participant is

the RT. The RT null hypothesis says that there is no effect of the intervention: The responses of the participant are independent of the condition (i.e., 'baseline' versus 'intervention') under which they are observed. We formulated the alternative hypothesis in a non-directional, two-tailed manner: There is a difference in self-esteem between the intervention phase and the baseline phase. In accordance with the alternative hypothesis, we used the absolute difference between the condition means as the test statistic for the RT (i.e., $T = |\bar{A} - \bar{B}|$). Note that this test statistic is analogous to the test statistic we used for the OLS analysis.

For our example, the observed test statistic (i.e., the value of the test statistic for the collected data) is 4.03. Constructing the randomization distribution implies that all 30 observed scores are kept fixed, whereas the start of the intervention phase is randomly determined taking into account the minimum of three measurement occasions for each phase. Accordingly, the test statistic can be calculated for each of the 25 assignments. Afterwards, the 25 test statistics can be sorted in ascending order, which forms the randomization distribution under the null hypothesis. The 25 test statistics range from 1.62 to 4.03. No other test statistic is as high as or higher than 4.03, the observed test statistic. Accordingly, the proportion of test statistics in the randomization distribution that exceeds or equals the observed test statistic is 1/25 or .04. Note that this is the smallest possible *p* value for this randomized AB design with 30 measurement occasions and a minimum of three measurement occasions per phase. We can reject the null hypothesis, and accept the alternative hypothesis, because the *p* value of the RT is smaller than the significance level α (.05 for our example). We conclude that that there is a statistically significant difference in self-esteem between the intervention phase and the baseline phase for this participant.

The RT analyses can be conducted using a free software package in R that assists researchers in designing and analyzing SCEDs using RTs: the SCDA package (Bulté &

Onghena, 2008, 2009, 2013). Appendix B includes the R code for testing the intervention effect for this dataset.

Empirical illustration: Testing the intervention effect for multiple participants within a replicated randomized AB design

We will now build further on the randomized AB design aimed at increasing the self-esteem of one student: In order to increase the external validity, it is possible that the experimenter planned to consecutively replicate the same experiment over three other students with low self-esteem as well. Accordingly, he would conduct a randomized sequential AB replication design (also called: replicated randomized AB design) with four participants. In accordance with the *What Works Clearinghouse* guidelines (Kratochwill et al., 2010), the experimenter conducted for each participant a randomized AB design with minimally three observations per phase. The selected assignments and collected data for each participant included in the example are included in Table 1 and plotted in Figure 2.

Using two-level HLM for testing the intervention effect for the replicated randomized AB design with four participants. We used OLS analysis to parametrically study the data relating to the first participant. For testing the intervention effect for the SCED with four participants, we could use OLS analysis to study the data relating to each included participant separately and estimate the case-specific treatment effect for each participant. Afterwards, these four treatment effects could be averaged and the between- and within-case variability could be calculated. However, a more efficient alternative is to use two-level HLM proposed by Van den Noortgate and Onghena (2003a, 2003b). The basic two-level HLM as well as extensions of the model are described in detail by Moeyaert, Ferron, Beretvas, and Van den Noortgate (2014). In the basic two-level HLM, measurement occasions (i) are situated at the first level, and go from 1 up to I. Measurement occasions are nested within a

case (j), that is situated at the second level. The cases included in one SCED study take on values from 1 to J. At the first level, the following regression equation can be used:

Eq. (2): $y_{ij} = \beta_{0j} + \beta_{1j} Phase_{ij} + e_{ij} \qquad e_{ij} \sim N(0,\sigma_e^2)$

The terms included in this regression equation are parallel to the ones discussed above for the OLS analysis. The only difference is that in this second equation the included cases are depicted by the symbol j. For instance, y_{ij} is the outcome score (i.e., self-esteem) for case j at measurement occasion i in the second regression equation. The model poses strong assumptions: It assumes that the within-case residuals (e_{ij}) are independently, identically, and normally distributed.

At the second level, the following regression equations can be used:

Eq. (3): $\begin{cases} \beta_{0j} = \theta_{00} + u_{0j} \\ \beta_{1j} = \theta_{10} + u_{1j} \end{cases} \qquad \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{u_0}^2 & \sigma_{u_0 u_1} \\ \sigma_{u_1 u_0} & \sigma_{u_1}^2 \end{bmatrix} \right)$

Because it is unlikely that the estimated baseline level and the treatment effect are identical for all participants within a study, the case-specific intercept (β_{0j}) and treatment effect (β_{1j}) from the first level regression equation (i.e., equation 2) are modeled to vary across the included participants in the regression equation relating to the second level (Eq. (3)). In the third regression equation, θ_{00} depicts the average baseline level and θ_{10} depicts the average treatment effect across the included participants. The equation includes participant-specific residuals (u_{0j} and u_{1j}), because each individual participant can have a baseline level and a treatment effect that deviate from the average baseline level (θ_{00}) and the average treatment effect (θ_{10}). Parallel to the first level residuals (Eq. (2)), the second level residuals (Eq. (3)) are assumed to be independently, identically, and multivariate normally distributed. The average baseline level (θ_{00}) can be used as an indicator for the need for the intervention: If the level of self-esteem is already high under the baseline condition, it might not be needed to intervene. The average treatment effect (θ_{10}) indicates the estimated

magnitude of the shift in the dependent variable (i.e., self-esteem) that tends to occur with the intervention. In addition to estimating the average baseline level (θ_{00}) and average treatment effect (θ_{10}), the two-level HLM can be used to estimate the between-case variance in the baseline level ($\sigma^2_{u_0}$) and the between-case variance in the treatment effect ($\sigma^2_{u_1}$), as well as the covariance between the baseline level and treatment effect ($\sigma_{u_0 u_1)}$. Single-case researchers using the two-level HLM are usually primarily interested in the average treatment effect over the included cases (θ_{10}). The null hypothesis that is tested is that the average treatment effect over all cases included in the study is zero.

Using the basic two-level HLM to analyze the data for the replicated randomized AB design with four participants (cf. Table 1) in SAS® 9.3, the estimated average baseline level across the participants ($\hat{\theta}_{00}$) was 2.31 (*SE* = 0.44), $t(2.95) = 5.21$, $p = .014$. The estimated average treatment effect across the participants ($\hat{\theta}_{10}$) was 3.26 (*SE* = 0.86), $t(3.01) = 3.80$, $p = .032$, indicating an increase in self-esteem due to the treatment that is statistically significant at an alpha level of .05. Appendix C includes the SAS code for analyzing this dataset.

Using RTcombiP for testing the intervention effect for the replicated randomized AB design with four participants. We used the RT to nonparametrically study the data related to the first participant. A nonparametric approach that can be used to test the intervention effect for the replicated randomized AB design with four participants is combining the *p* values from the RTs (i.e., RTcombiP; Edgington & Onghena, 2007). We will use the additive method for determining the significance of a sum of *p* values: Because the replicated randomized AB designs in our example provided independent tests of the same null hypothesis (i.e., the responses of the participant are independent of the condition under which they are observed), the *p* values of the RTs could be combined by first calculating the sum of the *p* values (S_{obs}) and then comparing this sum to all other sums *S* that could arise under the general null hypothesis (i.e., if the null hypothesis is true, then the *p* value is just a random

draw from a uniform [0,1] distribution; Onghena & Edgington, 2005). In contrast to HLM, the overall null hypothesis of RTcombiP is that there is no treatment effect for any of the cases included in the study. The combined *p* value is the proportion of combinations of *p* values which would give a sum *S* as small as the observed sum S_{obs}: $P(S \leq S_{obs}) = \sum_{k=0}^{\tilde{S}}(-1)^k\binom{n}{k}\frac{(S_{obs}-k)^n}{n!}$, with *n* = the number of *p* values to be combined, and with *k* = a counter up to the largest integer smaller than the observed sum $\tilde{S}$ = max ($k < S_{obs}$) (Onghena & Edgington, 2005).

The *p* values for the RTs, calculated based on the observed scores that are included in Table 1 (using the R code described in Appendix B), are respectively .04, .04, .04, and .12 for the four participants included in our replicated randomized AB design. Applying the formula above yields for S_{obs} = .04 + .04 + .04 + .12 = 0.24, and the largest integer smaller than 0.24, $\tilde{S}$ = 0, a combined *p* value of: $P(S \leq 0.24) = \sum_{k=0}^{0}(-1)^k\binom{4}{k}\frac{(0.24-k)^4}{4!} = \binom{4}{0}\frac{(0.24)^4}{4!}$ = .00014. Because the combined *p* value is statistically significant at the .05 level, we reject the null hypothesis that there is no treatment effect for any of the cases included in the study. The additive combined *p* value can be calculated using the SCDA package (Bulté & Onghena, 2013). Appendix D includes the R code for testing the intervention effect.

Objectives of the Monte Carlo simulation study

Single-case researchers should start their data analysis with visually inspecting the SCE data relating to level, trend, variability, immediacy of the effect, overlap between phases, and consistency of data patterns across similar phases (see e.g., Bulté & Onghena, 2012; Kratochwill et al., 2010). Afterwards, they can statistically test the intervention effect studied in their SCED. The aim of our Monte Carlo simulation study is to evaluate the performance of two-level HLM and RTcombiP for testing the intervention effect in replicated randomized AB designs. Our study hence aims to guide single-case researchers in their choice of an

appropriate method for statistically testing the intervention effect. More specifically, we studied the two-level HLM approach of Van den Noortgate and Onghena (2003a, 2003b) and the RTcombiP approach of Edgington and Onghena (2007). Although these two approaches are different in terms of their background, assumptions, functions, use, and capabilities in analyzing SCED data (cf. supra), both can be used to answer the question whether a certain intervention has a statistically significant effect on an outcome variable of interest when a replicated randomized AB design has been applied. We decided to use the HLM and the RTcombiP approach in our simulation study because of two reasons. First, they are currently considered two most promising inferential approaches for testing an intervention effect, that are often used for analyzing SCED data (see e.g., Huo, Heyvaert, Van Den Noortgate, & Onghena, 2014; Jenson, Clark, Kircher, & Kristjansson, 2007; Kratochwill & Levin, 2014; Shadish, Rindskopf, & Hedges, 2008). Second, statistical software is readily available for the two approaches (see Appendix A to D). This is interesting for us, in order to conduct our simulation study, but also to the applied single-case researcher, who can use the software for analyzing the collected SCED data.

Evaluating the performance of HLM and RTcombiP for testing the intervention effect in replicated randomized AB designs, we will focus on Type I error rate control and statistical power. Knowledge of Type I error rates and statistical power are critical for accurate application of any statistical test: A method with low statistical power will often fail to detect real effects that exist in the population, whereas a method with Type I error rates that exceed nominal rates (e.g., larger than 5% for nominal alpha = .05) does not control the risk of finding nonexistent effects (MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002, p. 2). Four factors will be manipulated in the simulation study: the mean intervention effect, the number of cases included in an SCED study, the number of measurement occasions for each case within one study, and between-case variance in the baseline level and in the treatment

effect. These factors were selected based on previous simulation studies on HLM and RTcombiP (Ferron, & Onghena, 1996; Ferron & Sentovich, 2002; Ferron et al., 2009; Levin, Ferron, & Kratochwill, 2012; Moeyaert, Ugille, Ferron, Beretvas, & Van den Noortgate, 2013a, 2013b; Ugille, Moeyaert, Beretvas, Ferron, & Van den Noortgate, 2012). The present simulation study is the first that evaluated the performance of HLM alongside RTcombiP. Accordingly, we did not *a priori* hypothesize whether Type I error rate control would be better for HLM or RTcombiP, and whether statistical power would be larger for HLM or RTcombiP.

Methods of the Monte Carlo simulation study

In the simulation study, we evaluated: (1) the *p* value of the estimated intervention effect across cases in a two-level HLM (using REML and using the Wald test), and (2) the *p* value obtained by combining RT *p* values using the additive method (with absolute difference between the phase means as the test statistic for the RT). The data were generated and analyzed using SAS® 9.3 Software (SAS Institute Inc., 2011-2015).

We simulated raw data for replicated randomized AB designs. In accordance with the *Design and Evidence Standards* for SCEDs developed by the *What Works Clearinghouse* (Kratochwill et al., 2010), we simulated the data in such a way that each phase at least consisted of three data points. The data were sampled from a normal distribution. Two factors were kept constant: The mean baseline level was 0 and the within-case variance was 1. Four factors were manipulated: (1) The mean intervention effect was 0 (i.e., no effect) or 2; (2) the number of cases included in a study was 3, 4, 5, 6, or 7; (3) the number of measurement occasions for each case within one study was 10, 20, 30, or 40 (this number was kept constant for all the cases included in one study); and (4) the between-case variance in the baseline level and in the treatment effect was 0, 0.1, 0.3, 0.5, 2, 4, 6, or 8. These parameter values for the

four factors were based on published meta-analyses of SCEDs (Denis, Van den Noortgate, & Maes, 2011; Heyvaert, Maes, Van den Noortgate, Kuppens, & Onghena, 2012; Heyvaert, Saenen, Maes, & Onghena, 2014; Kokina & Kern, 2010; Van den Noortgate & Onghena, 2008; Wang, Cui, & Parrila, 2011) and on analyses of characteristics of 809 published SCED studies (Shadish & Sullivan, 2011). We added a condition for zero between-case variance and two smaller values (0.1, 0.3) than usually observed in meta-analyses of SCEDs (Moeyaert et al., 2013a, 2013b; Ugille et al., 2012) to more closely study the effect of the between-case variance on Type I error rate control and statistical power for HLM and RTcombiP. Crossing the levels of these four factors led to a 2x5x4x8 factorial design, yielding 320 experimental conditions. In order to minimize simulation error, 10,000 data sets were simulated for each experimental condition. This resulted in a total of 3,200,000 data sets to analyze.

Each dataset was analyzed in two ways. First, we applied the two-level HLM approach of Van den Noortgate and Onghena (2003a, 2003b; cf. supra). We used the REML approach in SAS PROC MIXED to estimate the overall intervention effect (Littell et al., 2006). Furthermore, we used the Satterthwaite approach to approximate the degrees of freedom and to derive the corresponding *p* value, because this approach showed to provide accurate confidence intervals for the estimates of the average treatment effect for the two-level analysis of SCED data (Ferron et al., 2009). For HLM, we were primarily interested in the average treatment effect over the included cases. The null hypothesis that was tested by the two-level HLM analysis was that the average treatment effect over all cases included in the SCED study was zero.

Second, we applied the RTcombiP approach of Edgington and Onghena (2007; cf. supra). For RTcombiP we were interested in the *p* values of RTs combined using the additive method. We used the absolute difference between the condition means as the test statistic for the RT (i.e., $T=|\bar{A} - \bar{B}|$). For each case, the RT data analysis included: (1) Calculating the test

statistic for the 'observed' data (i.e., following the randomly selected assignment), (2) constructing the randomization distribution by looking at all possible assignments and calculating the test statistic for each of the assignments, and (3) determining the statistical significance of the observed test statistic by examining its position within the randomization distribution. Afterwards, the *p* values of the RT obtained for each case were combined for each study using the additive method. The overall null hypothesis that was tested by RTcombiP was that there was no treatment effect for any of the cases included in the SCED study. All tests were performed at the 5% significance level.

Results of the Monte Carlo simulation study

As a preliminary analysis to explore the most important patterns in the results, we studied variation between conditions using analyses of variance (ANOVAs) for HLM and RTcombiP with regard to Type I error rate and statistical power. We were especially interested in the proportion of variance explained by each of our parameters of interest: number of cases, number of measurement occasions, and between-case variance. Type I error rate for HLM was particularly explained by the between-case variance (R^2 = .727), followed by the number of cases included (R^2 = .102), and the number of measurement occasions for each case (R^2 = .035). Likewise, statistical power for HLM was particularly explained by the between-case variance (R^2 = .850), followed by the number of cases included (R^2 = .109), and the number of measurement occasions for each case (R^2 = .001).

Type I error rate for RTcombiP was particularly explained by the number of measurement occasions for each case (R^2 = .924), followed by the number of cases included (R^2 = .040), and the between-case variance (R^2 = .001). Likewise, statistical power for RTcombiP was particularly explained by the number of measurement occasions for each case

(R^2 = .875), followed by the number of cases included (R^2 = .060), and the between-case variance (R^2 = .039).

In what follows, the impact of the mean intervention effect, number of cases, number of measurement occasions, and between-case variance is described in detail for HLM and RTcombiP with regard to Type I error rate and statistical power. The results are presented in tabular format (Tables 2 to 5). Some trends for 3, 5, and 7 cases are illustrated in graphical format (Figures 3 to 6; trends for 4 and 6 cases are similar).

When the simulated mean intervention effect is 0 and the significance level is .05, we want the Type I error rate not to exceed .05 and to be as close as possible to .05. For HLM the Type I error rate was smaller than .05 in most conditions. Only for conditions with three cases and small between-case variance the Type I error rate was larger than .05. The Type I error rates for HLM are shown in Table 2 and Figure 3. For RTcombiP the Type I error rate was smaller than .05 in all conditions. Especially for the 10 data points conditions, the Type I error rate was far below .05 (ranging from .002 to .006). Type I error rates for RTcombiP are shown in Table 3 and Figure 4.

When the simulated mean intervention effect is different from zero (i.e., 2), we want the statistical power when testing the existence of an effect to be as high as possible. Cohen (1988) recommended statistical power values of .80 or higher. Overall, we saw that power for HLM was particularly dependent on the between-case variance (with larger between-case variance resulting in a substantial reduction of power), while power for RTcombiP was particularly dependent on the number of data points for the included cases (with a smaller number of data points resulting in smaller power). For HLM as well as RTcombiP, a larger number of cases included in the SCED resulted in higher statistical power.

Statistical power for HLM is depicted in Table 4 and Figure 5. Power was smaller than .80 for all conditions with a between-case variance of 4, 6, and 8. For the conditions with a

between-case variance of 2 and with 7 included cases, power was larger than .80. However, for all other conditions with a between-case variance of 2 power was smaller than .80. For all conditions with a between-case variance of 0.5 and with 5, 6 or 7 included cases, as well as for the conditions with a between-case variance of 0.5, with 4 included cases, and at least 20 data points per case, power was larger than .80. For the conditions with a between-case variance of 0.5 and with 3 included cases, as well as for the condition with 4 included cases but only 10 data points for each case, power was smaller than .80. For a between-case variance of 0.3, power was larger than .80 for all conditions with 4 or more cases, but smaller than .80 for the conditions with 3 cases. For a between-case variance of 0.1, power was larger than .80 in all conditions except the one with 3 cases and 10 data points for each case. For all conditions with a between-case variance of zero, statistical power was larger than .90.

Statistical power for RTcombiP is depicted in Table 5 and Figure 6. Power was larger than .80 for all conditions with 40 data points for the included cases. For the 30 data points conditions, power was larger than .80 for all conditions with 4, 5, 6, or 7 cases included. For the 30 data points and 3 cases conditions, power was larger than .80 for between-case variances of 0, 0.1, 0.3, and 0.5, but smaller than .80 for between-case variances of 2, 4, 6, and 8. For the 20 data points conditions, power was larger than .80 for all conditions with 5, 6, or 7 cases included. For the 20 data points and 3 or 4 cases conditions, power was larger than .80 for between-case variances of 0.5 or lower, but smaller than .80 for between-case variances of 2 or higher. We particularly noticed a difference between the statistical power results for RTcombiP between the 20 and 10 data points conditions (cf. Figure 6). For all 10 data points conditions, power was smaller than .60.

Discussion

The present Monte Carlo simulation study evaluated the performance of the two-level HLM approach of Van den Noortgate and Onghena (2003a, 2003b) and the RTcombiP approach of Edgington and Onghena (2007) for testing the intervention effect in replicated randomized AB designs. We examined the performance of the two approaches under various conditions relating to the mean intervention effect, the number of cases included in a study, the number of measurement occasions for each case within one study, and the between-case variance in the baseline level and in the treatment effect. We focused on Type I error rate control and statistical power for the two approaches.

For the two-level HLM approach of Van den Noortgate and Onghena (2003a, 2003b) Type I error rate was smaller than .05 in most conditions. For RTcombiP Type I error rate was smaller than .05 in all conditions. This is reassuring for the researcher using SCEDs even in harsh circumstances with a small number of cases and a small number of data points for each case. It means that both HLM and RTcombiP provide a valid test of the overall intervention effect. If there is no overall intervention effect then both statistical tests are guaranteed to keep the actual error rate below the nominal significance level as determined by the researcher.

As we discussed previously, we operationalized the two-level HLM approach with a specific model and with a specific estimation method for the present simulation study (e.g., REML approach, Satterthwaite approach). However, other HLM approaches and extensions have been developed and used for testing the intervention effect in randomized sequential replication designs (cf. supra). We acknowledge that the way we have operationalized the two-level HLM approach may have impacted our simulation study's Type I error rate control results.

The RTcombiP approach of Edgington and Onghena (2007) proved to be very conservative for the conditions with a small number of observations. For example, with ten

observations for each case and a nominal significance level of .05, the actual Type I error rate was below .01 in all conditions. This means that the researcher who wants to control the Type I error rate at 5% is actually much more stringent than intended. Although this is not a problem for the validity of the test, there is a trade-off by suppressing the statistical power of the test. This result for RTcombiP comes as no surprise because with ten observations and a minimum of three observations in each phase, there are only five possible randomizations. Because the smallest possible *p* value is the inverse of the number of possible randomizations, this implies that for the condition with ten observations the smallest possible *p* value for an individual RT is .20. So the statistical power of an individual RT at the 5% significance level under these circumstances is by definition zero. It is only by combining the smallest *p* values that a combined *p* value smaller than .05 can be obtained. For example, three *p* values of .20 result in a combined *p* value of .036 using RTcombiP (Edgington & Onghena, 2007; see Appendix D).

In our *Introduction*, we stated that we used the RTcombiP approach of Edgington and Onghena (2007) for the present simulation study, but that other RT approaches have been developed for testing the intervention effect in replicated randomized AB designs, such as the Marascuilo and Busk (1988) approach that works with the raw data instead of the *p* values. Using other RT approaches would lead to other results for the simulation study. For example, if we consider the results from the present simulation study for the condition with four cases included in the SCED, ten measurement occasions for each case, and a between-case variance equal to zero, the Type I error rate is .006 for the RTcombiP approach of Edgington and Onghena (2007) (see Table 3). This is substantially smaller than the Type I error estimate of .048, which is reported in the Ferron and Sentovich (2002) simulation study that examined the Marascuilo-Busk RT for the same condition (i.e., four cases included in the SCED, ten measurement occasions for each case, five potential intervention points per case, zero

between-case variance, and zero autocorrelation), suggesting that the RTcombiP approach is much more conservative.

Statistical power for the two-level HLM approach of Van den Noortgate and Onghena (2003a, 2003b) was particularly dependent on the between-case variance: Larger between-case variance resulted in a substantial reduction of power. HLM tests the null hypothesis that the mean treatment effect over all cases included in the study is zero. It is harder for HLM to detect a mean treatment effect when the between-case variance is larger: The uncertainty about our mean effect estimate is large, because an additional case could show a quite different treatment effect and therefore could considerably change our estimate of the mean effect. Similar to our remark for Type I error rate control, the way we have operationalized the two-level HLM approach may have impacted our simulation study's statistical power results.

Statistical power for the RTcombiP approach of Edgington and Onghena (2007) was particularly dependent on the number of data points for the included cases: Smaller numbers of data points resulted in smaller power. As discussed above, statistical power is related to the actual significance level. If the actual significance level is low, statistical power of the test is suppressed. Because RTcombiP proved to be a very conservative test for the conditions with a small number of observations, it is no surprise that statistical power is smaller for these conditions.

Comparing our results for the RTcombiP approach of Edgington and Onghena (2007) to the results of Ferron and Sentovich (2002)'s simulation study for the Marascuilo-Busk RT, we see that the applied RT approach considerably influences statistical power results. For example, if we consider the results from the present simulation study for the condition with four cases included in the SCED, ten measurement occasions for each case, and zero between-case variance, the estimated statistical power is .369 for the RTcombiP approach of Edgington

and Onghena (2007) (see Table 5). However, based on the results of Ferron and Sentovich (2002)'s simulation study, the power estimate for the Marascuilo-Busk RT was .866 for the same condition.

The present study has several practical implications for single-case researchers who are primarily interested in the question whether a certain intervention has statistically significant effects on the outcome variable of interest. When single-case researchers use a replicated randomized AB design, they could use HLM as well as RTcombiP in order to answer this research question. We see three options for the use of HLM and RTcombiP for analyzing the replicated randomized AB design data: (1) Using HLM *or* RTcombiP for testing the intervention effect, (2) applying a sequential approach by using *both* RTcombiP and HLM, and (3) directly combining RTcombiP and HLM for analyzing the data.

In the first option, a single-case researcher might decide to use HLM *or* RTcombiP, based on the specific null hypothesis he wishes to test, power differences, and the willingness to make the underlying assumptions associated with HLM and RTcombiP. When single-case researchers want to test the null hypothesis that the average treatment effect over all cases included in the SCED study is zero, they are advised to use the HLM approach. When single-case researchers want to test the null hypothesis that there is no treatment effect for any of the cases included in the SCED study, they are advised to use the RTcombiP approach. When the included cases' data are rather homogeneous (i.e., between-case variance of 0.5) and at least 4 cases are included, HLM offers sufficient statistical power, at least for the effect sizes used in our study. For SCEDs with a between-case variance of 2, HLM only offers sufficient statistical power when at least 7 cases are included in the study. When the included cases' data are more heterogeneous (i.e., between-case variance of 4 or more), we discourage using HLM based on this simulation study. When there are at least 40 data points for the cases included in the SCED, RTcombiP always offers sufficient statistical power. When there are at

least 30 data points for each case, researchers can use RTcombiP for analyzing SCEDs with 4, 5, 6, or 7 cases included. When there are at least 20 data points for each case, researchers can use RTcombiP for analyzing SCEDs with 5, 6, or 7 cases included. We discourage using RTcombiP when there are 10 or less data points for each case. For HLM as well as RTcombiP, a larger number of cases included in the SCED results in higher statistical power.

Our simulation study implies that including four cases in a replicated randomized AB design may already be sufficient for testing the intervention effect by means of HLM and RTcombiP, when the between-case variance is low (i.e., 0.5 or less) for HLM, and when the number of data points for the included cases is large (i.e., 30 or more) for RTcombiP. Moreover, for RTcombiP a replicated randomized AB design study including three cases may already yield sufficient power, when there are at least 40 data points for each included case.

In the second option, a single-case researcher might decide to use *both* RTcombiP and HLM for analyzing the replicated randomized AB design data, applying a sequential approach. The researcher could first use the RTcombiP approach to test the general null hypothesis of no treatment effect for any of the cases. The rationale behind selecting the RTcombiP approach for the first stage would be that the researcher prefers to use a nonparametric test, that relies on less stringent assumptions than parametric procedures, and that the RTcombiP approach is rather easy and straightforward to use. After determining that the intervention of interest has indeed a statistically significant effect on the outcome variable, the researcher could in a second stage use the HLM approach for analyzing the replicated randomized AB design data in closer detail, and to obtain the parameter estimate of and further model the average treatment effect and individual treatment effects. In this way, the Type I error rate is under control, even if the assumptions underlying the HLM are not valid. HLM could for instance be used to estimate the following parameters: case-specific intercepts and treatment effects, the average baseline level over the included cases, the average

treatment effect over the included cases, and estimates for within- and between-case variance in the baseline level and in the treatment effect (cf. supra). Furthermore, the researcher could extend the basic two-level model, for instance in order to account for trends (linear and non-linear; Shadish et al., 2013; Van den Noortgate & Onghena, 2003b), autocorrelation (Van den Noorgate & Onghena, 2003a), unequal within-phase variances (Baek & Ferron, 2013), external events (Moeyaert et al., 2013b), or non-normal outcomes, such as counts (Shadish et al., 2013). These more accurate HLMs in turn may increase the Type I error rate control and statistical power of the model.

A third option could be to directly combine RTcombiP and HLM for analyzing replicated randomized AB design data. HLM has the advantage of being a flexible modelling approach for all complexities in the data, but has the disadvantage of relying on questionable assumptions in many messy single-case circumstances. RTcombiP has the advantage of making minimal assumptions, but has the disadvantage of usually relying on simple test statistics, such as differences between means. We could combine the advantages of both procedures if we use the flexible modelling approach of HLM, but determine the statistical significance by RTcombiP, instead of using the conventional statistical distributions of HLM to compute the *p* value. This amounts to performing an RT on iterated HLM analyses for repartitioned data, with the HLM parameter estimate of the intervention effect as a test statistic. In that way we have an estimate of the magnitude of the intervention effect, taking into account all other factors included in the model, combined with valuable information relating to the (non)randomness of that intervention effect, without making distributional assumptions (cf. Heyvaert & Onghena, 2014). Although this third option is likely to be computer-intensive for many realistic data sets and models, applications seem to be straightforward using a smart RT wrapper (cf. Cassell, 2002). We think it is an interesting

venue for future research to address the operating characteristics of this kind of combination of HLM and RTcombiP.

There are several strengths related to the set-up of the present simulation study. First, the parameter values for the four factors that were manipulated for the simulation study were based on published meta-analyses of SCEDs and on analyses of characteristics of published SCED studies (cf. supra). Accordingly, the simulation study focused on realistic conditions. Second, the simulations were set up such that the modeling assumptions of the two-level HLM approach of Van den Noortgate and Onghena (2003a, 2003b) were accurate: The model used to simulate the data matched the model used to analyze the data. Third, the simulations were set up such that the use of the RTcombiP approach of Edgington and Onghena (2007) was statistically valid: The simulated SCEDs used randomization and the probabilities were computed based on randomization distributions that reflected the random assignment.

However, relating to the second point, one might question whether this aspect of the simulation study is representative for the conduct of 'real-life' SCEDs. For the simulation study we only modeled an effect of the treatment on the dependent variable. However, single-case researchers might worry about the possibility of factors other than the treatment impacting the time series, such as maturation, a historical effect, a testing effect, or instrumentation (cf. *Introduction*). If any of these effects are operating, regardless of whether the AB design has been replicated, it would suggest (unless specifically modeled, such as in Moeyaert et al., 2013b) some violation of the HLM assumptions. For instance, a recent study of Ferron, Moeyaert, Van den Noortgate, and Beretvas (2014) has shown that when an HLM is used in a context where there are other effects at play (like history or maturation), the Type I error rates can stray from the nominal value.

Relating to the third point as well, one might question whether this aspect of the simulation study is representative for the conduct of 'real-life' SCEDs: It might not always be

possible or desirable to include randomization in SCEDs (e.g., Kazdin, 1980). Past research has shown that when RTs are used in the absence of randomization, Type I error rates can stray from their nominal levels (e.g., Ferron, Foster-Johnson, & Kromrey, 2003; Levin et al., 2012; Manolov, Solanas, Bulté, & Onghena, 2010; Solanas, Sierra, Quera, & Manolov, 2008). However, stating that RTs and RTcombiP can *never* validly be used for nonrandomized SCEDs would be too restrictive too. The simulation studies of Ferron et al. (2003), Levin et al. (2012), Manolov et al. (2010), and Solanas et al. (2008) showed that there *are* non-randomized conditions where RTs maintained control of the Type I error rate. Furthermore, procedures have been developed that guarantee statistically valid RTs for SCED studies where randomizing is done *after* the study has begun (e.g., Edgington, 1975; Ferron & Jones, 2006; Ferron & Levin, 2014; Ferron & Ware, 1994; Koehler & Levin, 1998; Kratochwill & Levin, 2010), instead of doing it *a priori* (cf. *Introduction*). For instance, a single-case researcher can decide to make the manipulation of the conditions only partially dependent on the data: By using such a restricted random assignment, valid significance determination by the RT and RTcombiP approach is possible (Edgington, 1980). An example is a researcher conducting an AB design who *a priori* decides not to introduce the experimental treatment until the baseline data show stability (cf. response-guided experimentation), but to supplement this procedure with the random selection of the moment when the intervention phase starts, after baseline stability has been attained, thereby allowing for the valid use of an RT (Edgington, 1975).

The conclusions of this study are limited to the conditions that were simulated: The data were sampled from a normal distribution, there was no trend in the data, and the data were not autocorrelated. We also kept the number of measurement occasions constant for all the cases included in one study (i.e., 10, 20, 30, or 40 measurement occasions). However, these conditions may not be true for all 'real-life' SCEDs. For instance, it is possible that the

length of the data series differs between the cases included in one SCED study. Furthermore, it is possible that data in an SCED study are not normally distributed, that they are count data, binary data, or highly discrete data, or that there is trend in the data.

The present simulation study focused on testing the intervention effect for changes in level between the baseline and intervention conditions. When there are other effects of interest to the single-case researcher (e.g., changes in trend, changes in variance), HLM provides the opportunity to flexibly model these effects (cf. supra). Because we were for this simulation study interested in changes in level, we used the absolute difference between the phase means as the test statistic for the RT. When a single-case researcher expects or predicts other effects than changes in level, for instance changes in trend, another test statistic has to be chosen for the RT that corresponds to the expected effects. In our *Introduction*, we said that for the RTcombiP approach it is generally advised to select the test statistic in accordance with the kind of effects expected or predicted *a priori* to conducting the single-case experiment. However, if appropriate masking procedures are used (e.g., Ferron & Foster-Johnson, 1998; Ferron & Jones, 2006; Ferron & Levin, 2014) it is possible to guarantee a statistically valid RT when choosing a test statistic after the data have been collected.

Building on the present simulation study, future research can focus on conditions that might be more challenging for HLM and RTcombiP, such as a (linear) trend and autocorrelated data, and assess the impact of these conditions on Type I error rate control and statistical power for both approaches.

References

Baek, E., & 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.

Bulté, I., & Onghena, P. (2008). An R package for single-case randomization tests. *Behavior Research Methods, 40*(2), 467–478.

Bulté, I., & Onghena, P. (2009). Randomization tests for multiple baseline designs: An extension of the SCRT-R package. *Behavior Research Methods, 41*(2), 477–485.

Bulté, I., & Onghena, P. (2012). When the truth hits you between the eyes. A software tool for the visual analysis of single-case experimental data. *Methodology, 8*(3), 104–114.

Bulté, I., & Onghena, P. (2013). The single-case data analysis package: Analysing single-case experiments with R Software. *Journal of Modern Applied Statistical Methods, 12*(2), 28.

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

Center, B. J., Skiba, R. J., & Casey, A. (1985-1986). A methodology for the quantitative synthesis of intra-subject design research. *Journal of Special Education, 19*(4), 387–400.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Crosbie, J. (1993). Interrupted time-series analysis with brief single-subject data. *Journal of Consulting and Clinical Psychology, 61*(6), 966–974.

Crosbie, J. (1995). Interrupted time-series analysis with short series: Why it is problematic; how it can be improved. In J. M. Gottman (Ed.), *The analysis of change* (pp. 361–395). Mahwah, NJ: Erlbaum.

Denis, J., Van den Noortgate, W., & Maes, B. (2011). Self-injurious behavior in people with profound intellectual disabilities: A meta-analysis of single-case studies. *Research in Developmental Disabilities, 32*(3), 911–923.

Edgington, E. S. (1972). An additive method for combining probability values from independent experiments. *Journal of Psychology*, *80*(2), 351–363.

Edgington, E. S. (1975). Randomization tests for one-subject operant experiments. *Journal of Psychology*, *90*(1), 57–68.

Edgington, E. S. (1980). Overcoming obstacles to single-subject experimentation. *Journal of Educational Statistics, 5*(3), 261–267.

Edgington, E. S. (1996). Randomized single-subject experimental designs. *Behaviour Research and Therapy*, *34*(7), 567–574.

Edgington, E. S., & Onghena, P. (2007). *Randomization tests* (4th ed.). Boca Raton, FL: Chapman & Hall/CRC.

Egel, A. L., & Barthold, C. H. (2010). Single subject design and analysis. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 357–370). New York, NY: Routledge.

Ferron, J. M., & Foster-Johnson, L. (1998). Analyzing single-case data with visually guided randomization tests. *Behavior Research Methods, Instruments, & Computers, 30*(4), 698–706.

Ferron, J. M., & Jones, P. K. (2006). Tests for the visual analysis of response-guided multiple baseline data. *The Journal of Experimental Education, 75*(1), 66–81.

Ferron, J. M., & Levin, J. R. (2014). Single-case permutation and randomization statistical tests: Present status, promising new developments. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Statistical and methodological advances* (pp. 153–183). Washington, DC: American Psychological Association.

Ferron, J. M., & Onghena, P. (1996). The power of randomization tests for single-case phase designs. *The Journal of Experimental Education, 64*(3), 231–239.

Ferron, J. M., & Sentovich, C. (2002). Statistical power of randomization tests used with multiple-baseline designs. *The Journal of Experimental Education, 70*(2), 165–178.

Ferron, J. M., & Ware, W. (1994). Using randomization tests with responsive single-case designs. *Behavior Research and Therapy, 32*(7), 787–791.

Ferron, J. M., Bell, B. A., Hess, M. F., 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.

Ferron, J. M., Foster-Johnson, L., & Kromrey, J. (2003). The functioning of single-case randomization tests with and without random assignment. *The Journal of Experimental Education, 71*(3), 267–288.

Ferron, J. M., Moeyaert, M., Van den Noortgate, W., & Beretvas, S. N. (2014, June 16). Estimating causal effects from multiple-baseline studies: Implications for design and analysis. *Psychological Methods.* Advance online publication. doi:10.1037/a0037038

Heyvaert, M., & Onghena, P. (2014). Analysis of single-case data: Randomisation tests for measures of effect size. *Neuropsychological Rehabilitation, 24*(3-4), 507–527.

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*(2), 766–780.

Heyvaert, M., Saenen, L., Maes, B., & Onghena, P. (2014). Systematic review of restraint interventions for challenging behaviour among persons with intellectual disabilities: Focus on effectiveness in single-case experiments. *Journal of Applied Research in Intellectual Disabilities, 27*(6), 493–510.

Heyvaert, M., Wendt, O., Van den Noortgate, W., & Onghena, P. (in press). Randomization and data-analysis items in quality standards for single-case experimental studies. *Journal of Special Education.* doi:10.1177/0022466914525239

Holden, G., Bearison, D. J., Rode, D. C., Kapiloff, M. F., Rosenberg, G., & Onghena, P. (2003). Pediatric pain and anxiety: A meta-analysis of outcomes for a behavioral telehealth intervention. *Research on Social Work Practice, 13*(6), 675–692.

Huitema, B. E., & McKean, J. W. (1998). Irrelevant autocorrelation in least-squares intervention models. *Psychological Methods, 3*(1), 104–116.

Huo, M., Heyvaert, M., Van Den Noortgate, W., & Onghena, P. (2014). Permutation tests in the educational and behavioral sciences: A systematic review. *Methodology - European Journal of Research Methods for the Behavioral and Social Sciences, 10*(2), 43–59.

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*(5), 483–493.

Kazdin, A. E. (1980). Obstacles in using randomization tests in single-case experimentation. *Journal of Educational Statistics, 5*(3), 253–260.

Koehler, M. J., & Levin, J. R. (1998). Regulated randomization: A potentially sharper analytical tool for the multiple-baseline design. *Psychological Methods, 3*(2), 206–217.

Kokina, A., & Kern, L. (2010). Social story interventions for students with autism spectrum disorders: A meta-analysis. *Journal of Autism and Developmental Disorders, 40*(7), 812–826.

Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods, 15*(2), 124–144.

Kratochwill, T. R., & Levin, J. R. (Eds.). (2014). *Single-case intervention research: Statistical and methodological advances*. Washington, DC: American Psychological Association.

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 http://ies.ed.gov/ncee/wwc/documentsum.aspx?sid=229

Levin, J. R., & Wampold, B. E. (1999). Generalized single-case randomization tests: Flexible analyses for a variety of situations. *School Psychology Quarterly, 14*(1), 59–93.

Levin, J. R., Ferron, J. M., & Kratochwill, T. R. (2012). Nonparametric statistical tests for single-case systematic and randomized ABAB… AB and alternating treatment intervention designs: New developments, new directions. *Journal of School Psychology, 50*(5), 599–624.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). *SAS© system for mixed models* (2nd ed.). Cary, NC: SAS Institute Inc.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods, 7*(1), 83.

Maggin, D. M., Swaminathan, H., Rogers, H. J., O'Keeffe, B. V., Sugai, G., & Horner, R. H. (2011). A generalized least squares regression approach for computing effect sizes in

single-case research: Application examples. *Journal of School Psychology, 49*(3), 301–321.

Manolov, R., Solanas, A., Bulté, I., & Onghena, P. (2010). Data-division–specific robustness and power of randomization tests for ABAB designs. *The Journal of Experimental Education, 78*(2), 191–214.

Marascuilo, L. A., & Busk, P. L. (1988). Combining statistics for multiple-baseline AB and replicated ABAB designs across subjects. *Behavioral Assessment, 10*(1), 1–28.

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.

Moeyaert, M., Ugille, M., Ferron, J. M., Beretvas, S. N., & Van den Noortgate, W. (2013a). The three-level synthesis of standardized single-subject experimental data: A Monte Carlo simulation study. *Multivariate Behavioral Research, 48*(5), 719–748.

Moeyaert, M., Ugille, M., Ferron, J. M., Beretvas, S. N., & Van den Noortgate, W. (2013b). Modeling external events in the three-level analysis of multiple-baseline across-participants designs: A simulation study. *Behavior Research Methods, 45*(2), 547–559.

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. *The Journal of Experimental Education, 82*(1), 1–21.

O'Neill, B., & Findlay, G. (2014). Single case methodology in neurobehavioural rehabilitation: Preliminary findings on biofeedback in the treatment of challenging behaviour. *Neuropsychological Rehabilitation, 24*(3-4), 365–381.

Onghena, P., & Edgington, E. S. (2005). Customization of pain treatments: Single-case design and analysis. *Clinical Journal of Pain, 21*(1), 56–68.

SAS Institute Inc. (2011-2015). *SAS® 9.3 Software* [Computer software].

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.

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*(3), 385–405.

Shadish, W. R., Rindskopf, D. M., & Hedges, L. V. (2008). The state of the science in the meta-analysis of single-case experimental designs. *Evidence-Based Communication Assessment and Intervention, 2*(3), 188–196.

Solanas, A., Sierra, V., Quera, V., & Manolov, R. (2008). Random assignment of intervention points in two-phase single-case designs: Data-division-specific distributions. *Psychological Reports, 103*(2), 499–515.

ter Kuile, M. M., Bulté, I., Weijenborg, P. T., Beekman, A., Melles, R., & Onghena, P. (2009). Therapist-aided exposure for women with lifelong vaginismus: A replicated single-case design. *Journal of Consulting and Clinical Psychology, 77*(1), 149–159.

Todman, J. B., & Dugard, P. (2001). *Single-case and small-n experimental designs: A practical guide to randomization tests*. Mahwah, NJ: Erlbaum.

Ugille, M., Moeyaert, M., Beretvas, S. N., Ferron, J. M., & Van den Noortgate, W. (2012). Multilevel meta-analysis of single-subject experimental designs: A simulation study. *Behavior Research Methods, 44*(4), 1244–1254.

Van de Vliet, P., Onghena, P., Knapen, J., Fox, K. R., Probst, M., van Coppenolle, H., & Pieters, G. (2003). Assessing the additional impact of fitness training in depressed psychiatric patients receiving multifaceted treatment: A replicated single-subject design. *Disability & Rehabilitation, 25*(24), 1344–1353.

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

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

Van den Noortgate, W., & Onghena, P. (2008). A multilevel meta-analysis of single-subject experimental design studies. *Evidence-based Communication Assessment and Intervention, 2*(3), 142–151.

Vlaeyen, J. W., de Jong, J., Geilen, M., Heuts, P. H., & van Breukelen, G. (2001). Graded exposure in vivo in the treatment of pain-related fear: A replicated single-case experimental design in four patients with chronic low back pain. *Behaviour Research and Therapy, 39*(2), 151–166.

Wampold, B. E., & Worsham, N. L. (1986). Randomization tests for multiple-baseline designs. *Behavioral Assessment, 8*, 135–143.

Wang, S., Cui, Y., & Parrila, R. (2011). Examining the effectiveness of peer-mediated and video-modeling social skills interventions for children with autism spectrum disorders: A meta-analysis in single-case research using HLM. *Research in Autism Spectrum Disorders, 5*(1), 562–569.

Table 1

Selected assignments and collected data for a randomized sequential replication design, including four participants: For each participant a randomized AB phase design was used, with minimally three observations per phase

Participant 1		Participant 2		Participant 3		Participant 4	
Phase	Data	Phase	Data	Phase	Data	Phase	Data
A	1	A	4	A	3	A	1
A	2	A	3	A	2	A	2
A	1	A	4	A	3	A	1
A	1	A	3	A	2	A	1
A	2	A	3	A	3	A	2
A	2	A	4	A	3	A	2
A	2	A	3	A	3	A	1
A	1	B	6	A	2	A	2
A	1	B	7	A	3	A	1
A	2	B	7	A	2	A	2
A	2	B	6	A	5	A	2
B	6	B	7	A	3	A	2
B	5	B	7	A	2	A	2
B	6	B	6	A	2	A	2
B	5	B	7	A	2	A	1
B	6	B	7	A	2	A	2
B	6	B	6	A	3	A	2
B	5	B	7	A	3	B	2
B	5	B	6	A	2	B	3
B	6	B	7	A	3	B	3
B	6	B	6	B	7	B	2
B	5	B	6	B	7	B	3
B	6	B	7	B	7	B	2
B	6	B	7	B	8	B	3
B	5	B	7	B	8	B	3
B	6	B	7	B	7	B	2
B	5	B	6	B	8	B	2
B	5	B	7	B	8	B	2
B	6	B	7	B	7	B	3
B	6	B	7	B	8	B	3

Table 2

Type I error rate for the HLM approach

Number of data points	Between-case variance	Number of cases included				
		3 cases	4 cases	5 cases	6 cases	7 cases
10	0	.034	.031	.032	.031	.031
	0.1	.050	.040	.042	.045	.045
	0.3	.055	.048	.047	.046	.043
	0.5	.054	.047	.043	.043	.042
	2	.041	.043	.035	.035	.034
	4	.036	.032	.033	.032	.033
	6	.038	.033	.032	.033	.032
	8	.036	.033	.035	.032	.032
20	0	.034	.029	.031	.028	.029
	0.1	.047	.049	.043	.043	.046
	0.3	.051	.046	.043	.039	.041
	0.5	.049	.044	.041	.038	.039
	2	.038	.036	.035	.030	.032
	4	.039	.033	.035	.029	.032
	6	.036	.030	.032	.032	.032
	8	.035	.028	.030	.029	.033
30	0	.030	.026	.028	.029	.028
	0.1	.055	.045	.048	.042	.042
	0.3	.048	.040	.041	.043	.038
	0.5	.046	.040	.038	.036	.037
	2	.035	.030	.032	.031	.032
	4	.031	.033	.032	.031	.037
	6	.031	.027	.029	.035	.032
	8	.035	.033	.031	.035	.032
40	0	.030	.028	.029	.032	.027
	0.1	.055	.044	.045	.039	.040
	0.3	.045	.043	.038	.037	.039
	0.5	.045	.037	.037	.036	.038
	2	.035	.033	.032	.032	.035
	4	.035	.031	.032	.035	.030
	6	.036	.035	.032	.029	.031
	8	.036	.030	.031	.029	.028

Table 3

Type I error rate for the RTcombiP approach

Number of data points	Between-case variance	Number of cases included				
		3 cases	4 cases	5 cases	6 cases	7 cases
10	0	.005	.006	.005	.003	.003
	0.1	.005	.006	.005	.004	.003
	0.3	.006	.005	.004	.004	.002
	0.5	.006	.005	.004	.003	.002
	2	.004	.005	.004	.004	.002
	4	.005	.006	.004	.004	.002
	6	.005	.005	.004	.003	.003
	8	.006	.006	.004	.004	.002
20	0	.028	.020	.020	.019	.015
	0.1	.028	.020	.019	.018	.018
	0.3	.025	.021	.018	.017	.018
	0.5	.028	.021	.021	.018	.016
	2	.027	.020	.019	.016	.016
	4	.030	.021	.017	.019	.016
	6	.029	.022	.019	.015	.016
	8	.028	.020	.019	.017	.015
30	0	.029	.033	.024	.026	.025
	0.1	.029	.031	.025	.024	.026
	0.3	.031	.028	.026	.025	.027
	0.5	.029	.030	.026	.025	.023
	2	.029	.032	.026	.025	.031
	4	.030	.031	.021	.023	.024
	6	.029	.031	.026	.024	.025
	8	.028	.031	.023	.028	.026
40	0	.036	.033	.030	.031	.030
	0.1	.032	.035	.035	.031	.032
	0.3	.036	.030	.032	.030	.026
	0.5	.038	.031	.034	.031	.030
	2	.034	.032	.033	.029	.031
	4	.031	.031	.029	.030	.029
	6	.036	.031	.034	.035	.031
	8	.037	.032	.032	.026	.030

Table 4

Statistical power for the HLM approach

Number of data points	Between-case variance	Number of cases included				
		3 cases	4 cases	5 cases	6 cases	7 cases
10	0	**.929**	**.994**	**.999**	**1**	**1**
	0.1	.793	**.974**	**.997**	**.999**	**1**
	0.3	.616	**.880**	**.977**	**.997**	**1**
	0.5	.524	.791	**.940**	**.984**	**.995**
	2	.264	.424	.577	.707	**.813**
	4	.171	.274	.373	.467	.565
	6	.135	.204	.263	.345	.428
	8	.117	.161	.223	.281	.337
20	0	**.955**	**.995**	**.999**	**1**	**1**
	0.1	**.839**	**.989**	**.999**	**1**	**1**
	0.3	.666	**.940**	**.994**	**1**	**1**
	0.5	.552	**.856**	**.966**	**.996**	**.999**
	2	.269	.453	.622	.743	**.840**
	4	.176	.284	.389	.490	.581
	6	.134	.203	.282	.350	.432
	8	.114	.167	.235	.290	.341
30	0	**.956**	**.996**	**.999**	**1**	**1**
	0.1	**.866**	**.993**	**.999**	**1**	**1**
	0.3	.692	**.958**	**.997**	**1**	**1**
	0.5	.576	**.880**	**.979**	**.997**	**1**
	2	.270	.457	.627	.756	**.852**
	4	.168	.281	.386	.491	.575
	6	.138	.205	.290	.357	.423
	8	.113	.167	.227	.291	.333
40	0	**.962**	**.997**	**1**	**1**	**1**
	0.1	**.889**	**.994**	**.999**	**1**	**1**
	0.3	.715	**.966**	**.998**	**1**	**1**
	0.5	.588	**.896**	**.983**	**.998**	**1**
	2	.268	.469	.631	.763	**.854**
	4	.176	.285	.384	.497	.580
	6	.139	.200	.283	.363	.431
	8	.120	.163	.223	.289	.347

Note: Values equal to or larger than .80 are tabulated in bold.

Table 5

Statistical power for the RTcombiP approach

Number of data points	Between-case variance	Number of cases included				
		3 cases	4 cases	5 cases	6 cases	7 cases
10	0	.235	.369	.468	.542	.600
	0.1	.225	.355	.451	.519	.586
	0.3	.221	.342	.423	.482	.544
	0.5	.213	.319	.395	.452	.501
	2	.207	.280	.325	.370	.414
	4	.224	.294	.347	.394	.424
	6	.250	.333	.372	.425	.472
	8	.279	.356	.416	.461	.507
20	0	**.921**	**.967**	**.989**	**.997**	**.999**
	0.1	**.908**	**.953**	**.984**	**.994**	**.998**
	0.3	**.856**	**.915**	**.962**	**.985**	**.993**
	0.5	**.813**	**.872**	**.932**	**.967**	**.981**
	2	.677	.722	**.826**	**.876**	**.911**
	4	.660	.728	**.811**	**.868**	**.903**
	6	.679	.737	**.832**	**.879**	**.923**
	8	.708	.755	**.858**	**.904**	**.928**
30	0	**.981**	**.998**	**1**	**1**	**1**
	0.1	**.973**	**.995**	**.999**	**1**	**1**
	0.3	**.939**	**.979**	**.994**	**.998**	**1**
	0.5	**.900**	**.954**	**.981**	**.993**	**.998**
	2	.756	**.848**	**.905**	**.946**	**.969**
	4	.740	**.837**	**.896**	**.929**	**.963**
	6	.751	**.840**	**.905**	**.939**	**.969**
	8	.768	**.852**	**.919**	**.949**	**.975**
40	0	**.997**	**1**	**1**	**1**	**1**
	0.1	**.994**	**.999**	**1**	**1**	**1**
	0.3	**.970**	**.990**	**.999**	**1**	**1**
	0.5	**.940**	**.976**	**.993**	**.998**	**.999**
	2	**.803**	**.880**	**.942**	**.968**	**.981**
	4	.792	**.871**	**.930**	**.959**	**.978**
	6	**.805**	**.880**	**.940**	**.969**	**.982**
	8	**.816**	**.895**	**.949**	**.970**	**.984**

Note: Values equal to or larger than .80 are tabulated in bold.

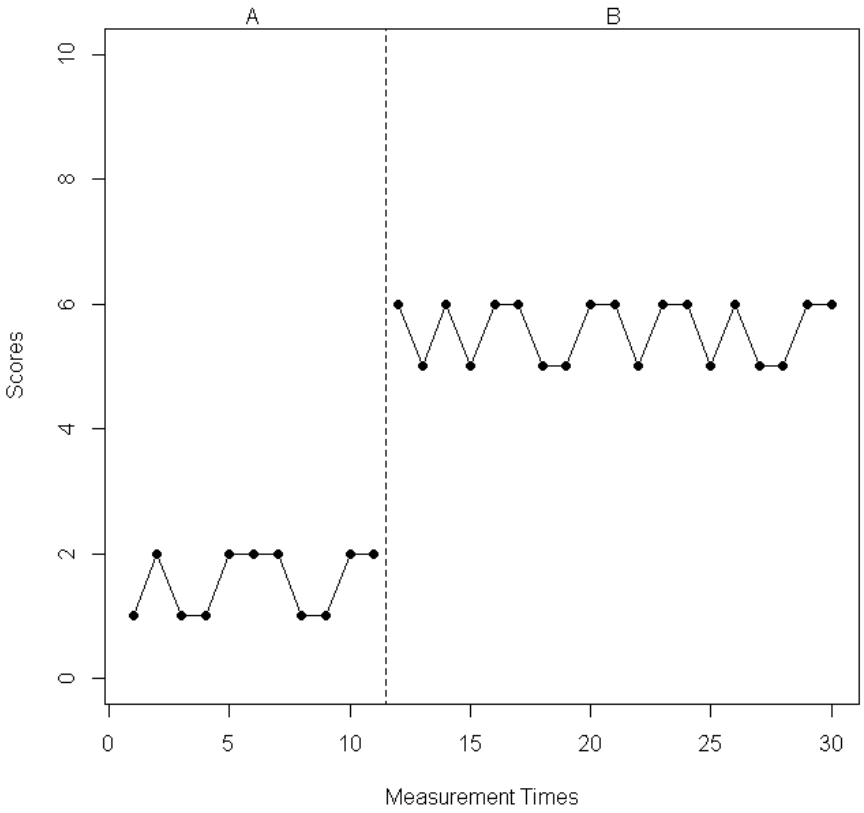
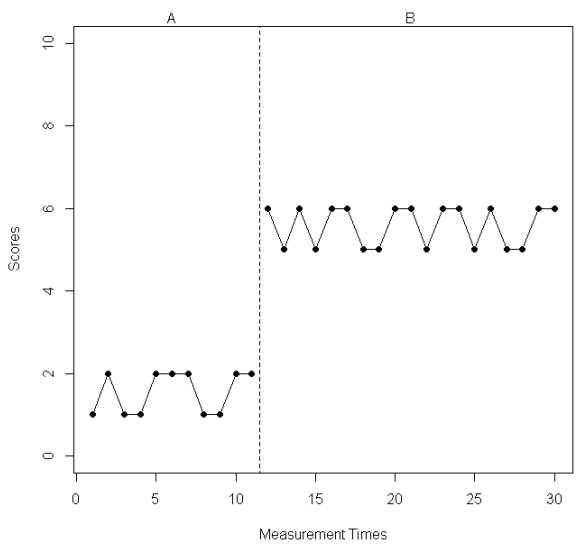
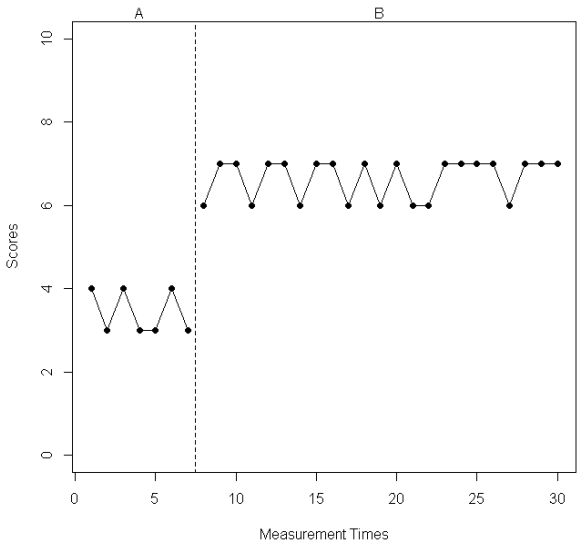


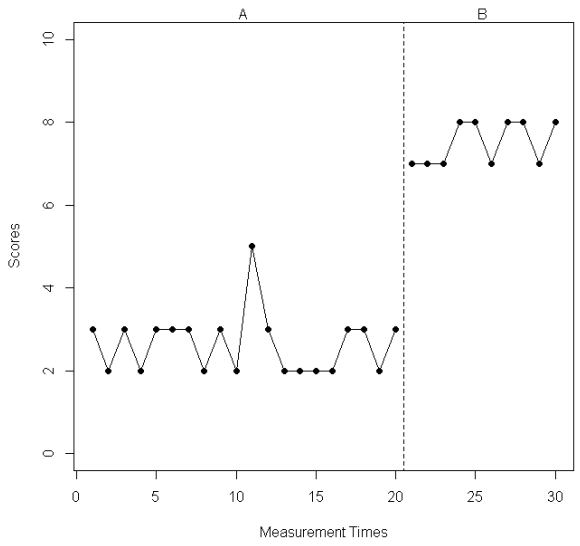
Figure 1. Example of a randomized AB phase design.



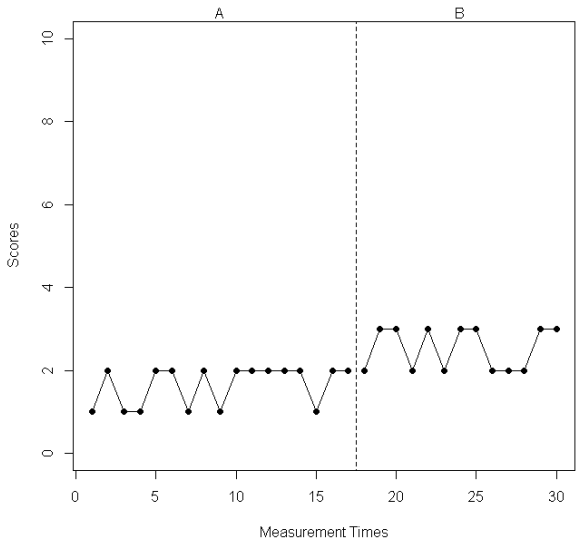
Participant 1



Participant 2



Participant 3



Participant 4

Figure 2. Example of a randomized sequential replication design including four participants.

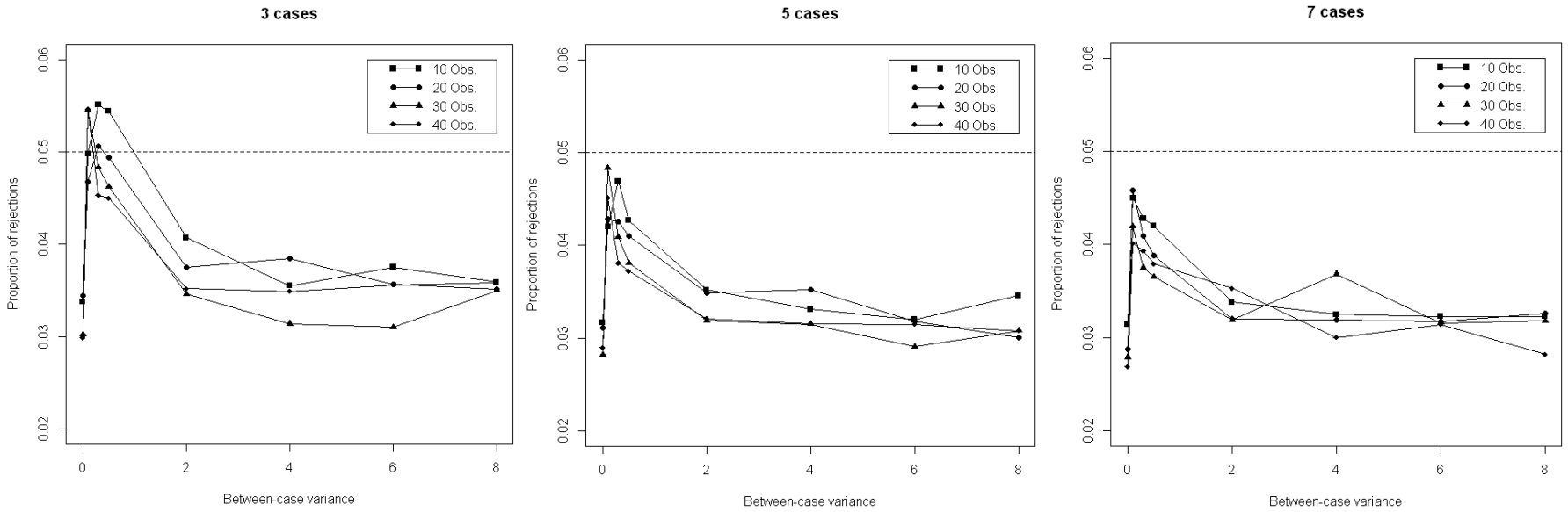


Figure 3. Type I error rates for the HLM approach for 3, 5, and 7 cases.

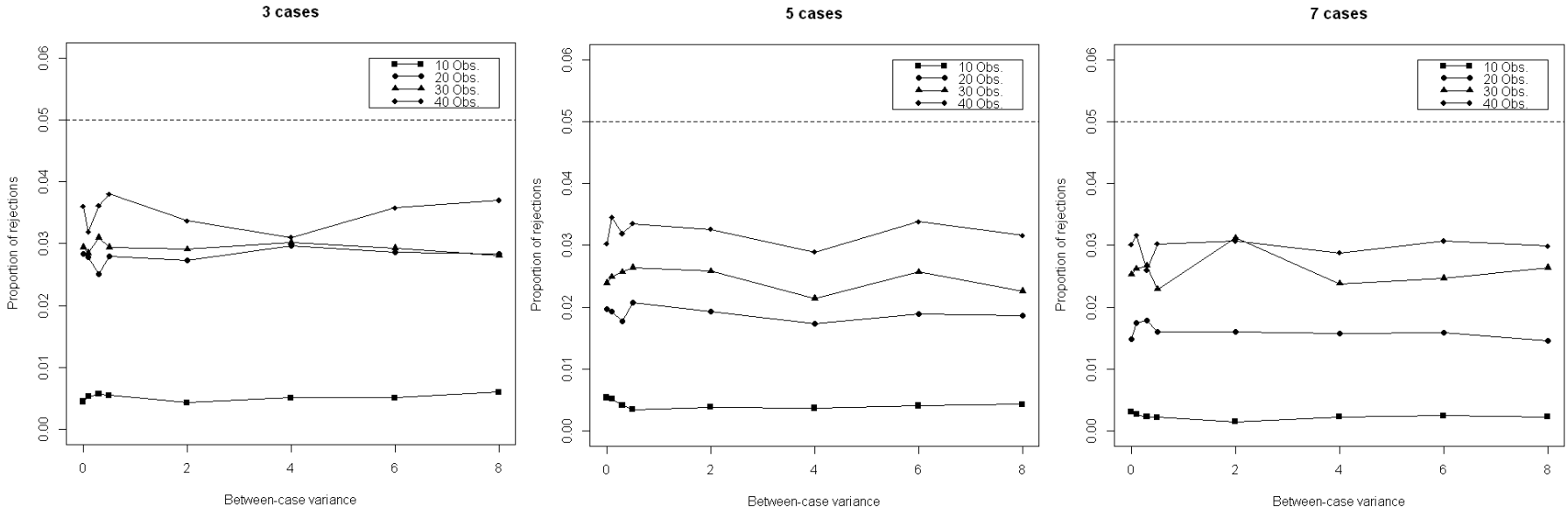


Figure 4. Type I error rates for the RTcombiP approach for 3, 5, and 7 cases.

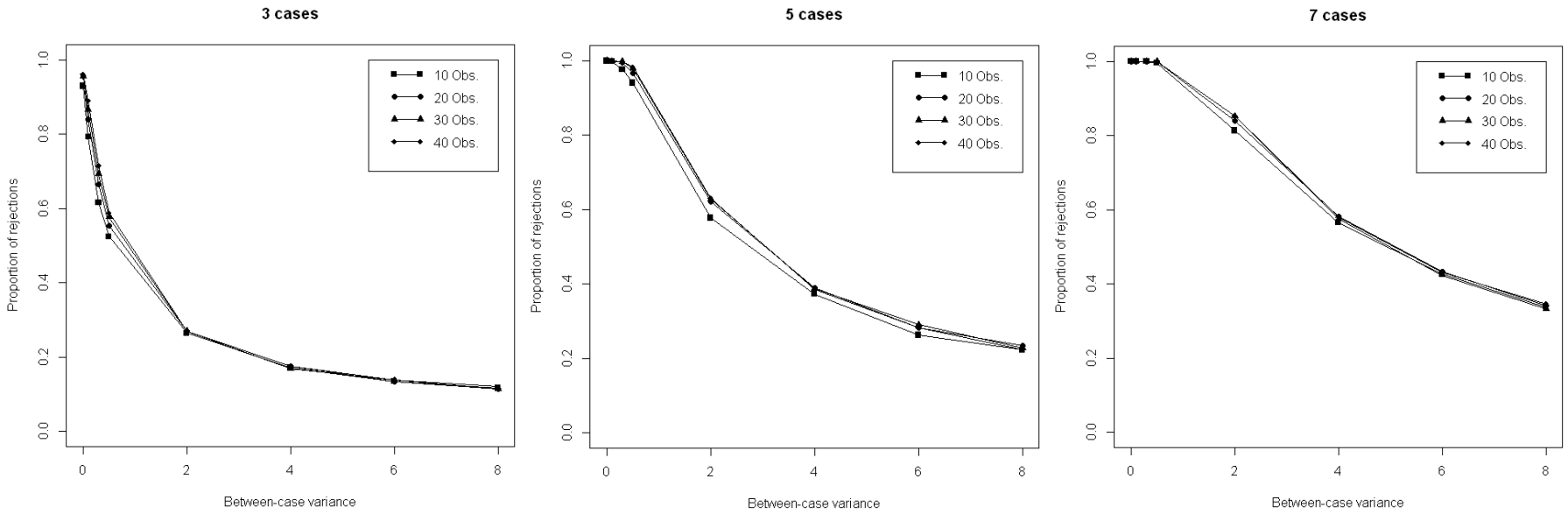


Figure 5. Statistical power for the HLM approach for 3, 5, and 7 cases.

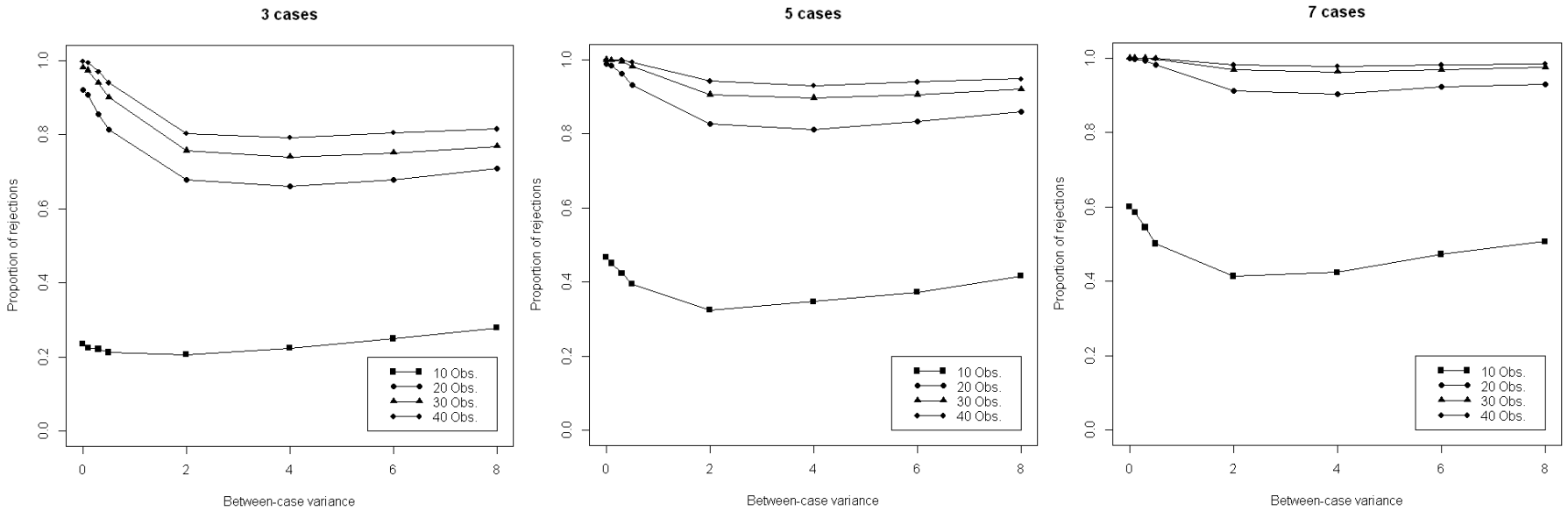


Figure 6. Statistical power for the RTcombiP approach for 3, 5, and 7 cases.

Appendix A: SAS code for analyzing the data from one participant within a randomized AB phase design using the parametric OLS approach

```
proc reg data= EXAMPLE;
model Y = Phase;
by case;
run;
```

Appendix B: R code for analyzing the data from one participant within a randomized AB phase design using the nonparametric RT approach included in the SCDA package (Bulté & Onghena, 2008, 2009, 2013)

```
library("SCRT")
quantity(design="AB",MT=30,limit=3)
observed(design="AB",statistic="|A-B|",data = read.table(EXAMPLE))
distribution.systematic(design="AB",statistic="|A-B|",save="no",limit=3,data =
read.table(EXAMPLE))
pvalue.systematic(design="AB",statistic="|A-B|",save="no",limit=3,data =
read.table(EXAMPLE))
```

Appendix C: SAS code for analyzing the data from multiple participants within a randomized sequential replication design using the parametric HLM approach

```
proc mixed data= EXAMPLE method=REML;
class case;
model Y = Phase / solution ddfm=sat;
random Intercept Phase / sub=Case;
run;
```

Appendix D: R code for analyzing the data from multiple participants within a randomized sequential replication design using the nonparametric RTcombiP approach included in the SCDA package (Bulté & Onghena, 2008, 2009, 2013)

```
library("SCMA")
combine("+", pvalues = read.table(EXAMPLE))
```