Perspectives on Behavior Science
https://doi.org/10.1007/s40614-021-00304-z

SI: ADVANCED QUANTITATIVE TECHNIQUES FOR SINGLE CASE EXPERIMENTAL DESIGN

# The Power to Explain Variability in Intervention Effectiveness in Single-Case Research Using Hierarchical Linear Modeling

Mariola Moeyaert[1] · Panpan Yang[1] · Xinyun Xu[1]

Accepted: 14 June 2021/


## Abstract

This study investigated the power of two-level hierarchical linear modeling (HLM) to explain variability in intervention effectiveness between participants in context of single-case experimental design (SCED) research. HLM is a flexible technique that allows the inclusion of participant characteristics (e.g., age, gender, and disability types) as moderators, and as such supplements visual analysis findings. First, this study empirically investigated the power to estimate intervention and moderator effects using Monte Carlo simulation techniques. The results indicate that larger values for the true effects and the number of participants resulted in a higher power. The more moderators added to the model, the more participants needed to detect the effects with sufficient power (i.e., power $\geq$.80). When a model includes three moderators, at least 20 participants are required to capture the intervention effect and moderator effects with sufficient power. For that same condition, but only including one moderator, seven participants are sufficient. Specific recommendations for designing a SCED study with sufficient power to estimate intervention and moderator effects were provided. Second, this study introduced a newly developed user-friendly point and click Shiny tool, *PowerSCED*. This tool assists applied SCED researchers in designing a SCED study that has sufficient power to detect intervention and moderator effects. To end, the use of HLM with the inclusion of moderators was demonstrated using two previously published SCED studies in the journal *School Psychology Quarterly*.

**Keywords** two-level hierarchical linear model · single-case experimental design · moderators · statistical power · *PowerSCED* Shiny tool

✉ Mariola Moeyaert
mmoeyaert@albany.edu

[1] School of Education, Department of Educational and Counseling Psychology, Division of Educational Psychology and Methodology, The University at Albany-SUNY, 1400 Washington Avenue, Albany, NY 12222, USA

In a single-case experimental design (SCED) study, individual cases (e.g., participants) are repeatedly measured during a baseline condition, followed by an intervention condition (Kazdin, 2011; Kratochwill et al., 2014). By comparing outcomes obtained during the baseline condition and the intervention condition, researchers can evaluate whether there is a basic effect of the intervention on the outcome variable (Kratochwill et al., 2010; What Works Clearinghouse, 2020). A basic effect is documented if the outcome pattern observed during the intervention condition is not a continuation of the pattern observed during the baseline condition. The pattern is traditionally investigated by visually analyzing the level, trend, and variability (Kratochwill et al., 2010). If evidence in support of a basic effect is demonstrated at least three times at three different time points, then the researcher has preliminary evidence to conclude that the change in the outcome measure(s) is caused by the intervention, not by outside experimental factors (Moeyaert et al., 2013a; What Works Clearinghouse, 2020). Several SCED types can be used to document this basic effect at least three times at three different moments in time. These designs can be classified as either within series SCEDs or between series SCEDs (Barlow et al., 2008; Horner & Odom, 2014). The within-series designs involve replication of the basic effect within one participant, whereas the between series designs involve replication of the basic effect across participants. The ABAB reversal designs (Ferron, 2005), alternating treatment designs (Wolery et al., 2018), and chancing criterion designs (Gast & Ledford, 2018) are examples of within series designs, and the multiple-baseline designs and the multiple-probe design (Gast et al., 2018) are between-series designs (Shadish & Sullivan, 2011; Jamshidi et al., 2020; Moeyaert et al., 2020).

Jamshidi et al. (2020) conducted a systematic review of SCED studies, focusing on SCED design characteristics. Their review indicated that 56% of the 177 included SCED studies used a multiple-baseline design (MBD). The MBD across participants embeds replication across participants, and thus is more externally valid (Ferron & Scott, 2005). Figure 1 displays an example of an MBD across six participants. As depicted in Figure 1, King et al. (2017) implemented the On-Task in a Box treatment for increasing on-task behavior of highly off-task students at the 4$^{th}$, 5$^{th}$, and 6$^{th}$ observation occasion for Participants 1 and 4, 2 and 5, 3 and 6, respectively. Hence, studies using MBDs can make more generalized conclusions about intervention effectiveness (i.e., demonstration of effectiveness across six participants, at three different points in time). In addition, the MBD introduces the intervention staggered across participants, and thus it is internally valid (Ferron & Scott, 2005). The staggered start of the intervention in MBDs prevents outside experimental factors as intervention confounders (i.e., the threat of history; Shadish et al., 2002). Using Figure 1 as an example, only Participants 1 and 4 are anticipated to experience a change in outcome pattern after the 4$^{th}$ measurement observation (when Participants 1 and 4 receive the intervention), whereas the outcome patterns of the other participants (still in the baseline condition) are expected to remain at baseline levels. The current study focused on the MBD because of its popularity and its high internal and external validity.

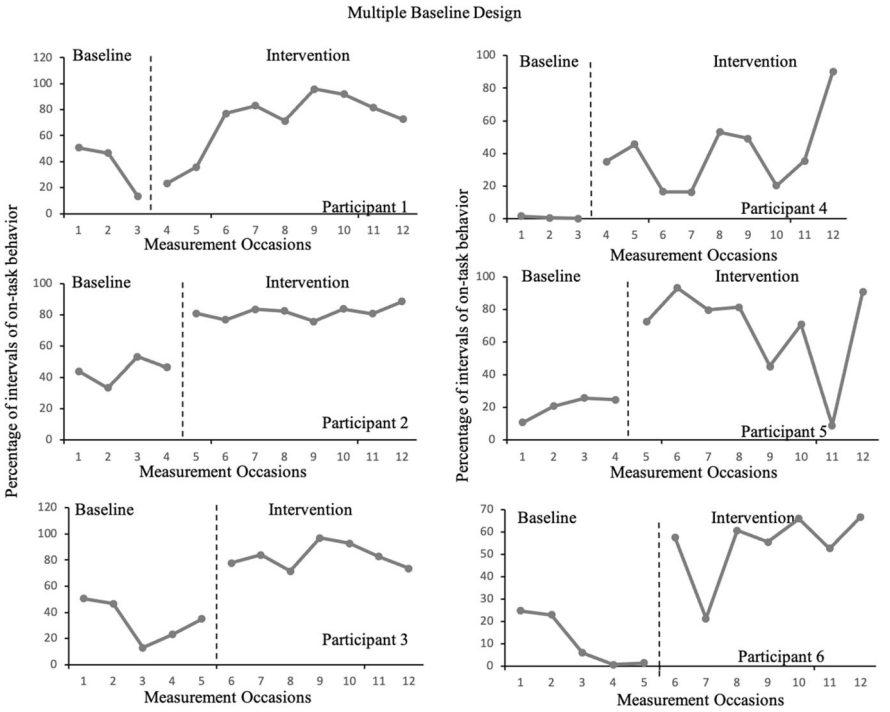


**Fig. 1** Multiple-Baseline Design Across Participants. *Note*: The figure was recreated using Microsoft® Excel. The raw data was extracted from King et al. (2017) using the data retrieval program WebPlotDigitizer

## Quantification of Single-Case Research Using Two-Level Hierarchical Linear Modeling

In addition to visually documenting evidence in support of intervention effectiveness (Barton et al., 2018; Kratochwill et al., 2014), multiple statistics have been developed to quantify the effect of the intervention in SCEDs. For instance, nonoverlapping statistics, such as NAP (Parker et al., 2009) and Tau indices (Parker et al., 2011; Tarlow, 2017), use the percentage of nonoverlapping outcome scores between the baseline condition and the intervention condition to quantify intervention effectiveness. Parametric statistics, such as regression-based effect sizes, quantify the magnitude of changes in outcome patterns (e.g., changes in outcome levels and/or outcome trends) between baseline and intervention conditions together with its statistical significance (Ferron et al., 2017; Moeyaert et al., 2014a; Van den Noortgate & Onghena, 2003a, 2003b). This study focused on this latter category as regression-based statistics allow for quantifying the magnitude of the intervention effect, and for evaluating whether the intervention effect is clinically and statistically significant. The outcome measure $Y_i$ can be regressed on a dummy coded variable, $D_i$ indicating the condition an observation belongs to (i.e., $D_i$ equals 0 when $Y_i$ belongs to the baseline condition, intervention condition otherwise). The within-participant errors ($e_i$) are assumed to be homogeneous, independent and normally distributed around an average of 0 and variance $\sigma_e^2$.

Therefore, $\sigma_e^2$ indicates the within-participant error variance. This regression model is displayed in Eq. 1.

$$Y_i = \beta_0 + \beta_1 D_i + e_i \quad \text{with } e_{ij} \sim N\left(0, \sigma_e^2\right) \tag{1}$$

Using Eq. 1, the intervention effect (i.e., $\widehat{\beta}_1$, change in outcome level between the baseline and the intervention conditions) can be estimated for each participant separately. To estimate the intervention effect across multiple participants (and even across multiple studies), the mean, weighted mean, median and/or range can be reported (Jamshidi et al., 2020). As an alternative, hierarchical linear modeling (HLM) can be used, and is recommended, because this approach considers the nested MBD structure (repeated measures are nested within participants). Given that the repeated measures of one participant are more alike compared to repeated measures of other participants, the HLM approach is preferred as it considers this dependency. Using HLM, the overall intervention effectiveness across participants is quantified, in addition to individual differences in intervention effectiveness (Van den Noortgate & Onghena, 2003a, 2003b). The two-level hierarchical linear model is a straightforward extension of the simple linear regression, and is displayed in Eq. 2. The first level represents the measurement level (within-participants). The dependent variable $Y_{ij}$ is the outcome of participant ($j$) at measurement occasion ($i$). The independent variable $D_{ij}$ is a dummy variable, indicating the condition. Therefore, $\beta_{0j}$ indicates the expected baseline level for participant $j$ and $\beta_{1j}$ indicates the expected intervention effect (i.e., change between baseline level and intervention level) for participant $j$.[1]

$$\begin{gathered}\text{Level 1 (within participants)}: Y_{ij} = \beta_{0j} + \beta_{1j} D_{ij} + e_{ij} \quad \text{with } e_{ij} \sim N\left(0, \sigma_e^2\right) \\ \text{Level 2 (across participants)}: \begin{cases} \beta_{0j} = \theta_0 + u_{0j} \\ \beta_{1j} = \theta_1 + u_{1j} \end{cases} \\ \text{with } \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim MVN\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{u_0}^2 & \\ \sigma_{u_0 u_1} & \sigma_{u_0}^2 \end{bmatrix} \right), \text{and } e_{ij} \sim N\left(0, \sigma_e^2\right)\end{gathered} \tag{2}$$

The level 1 parameters (at the right side of the equation sign) are allowed to vary at the second level as it is reasonable to expect that baseline levels ($\beta_{0j}'s$) and intervention effects ($\beta_{1j}'s$) vary between participants. Therefore, these parameters are a function of an overall average effect across participants (reflected by the $\theta$'s) and individual differences (reflected by the u's). $\theta_0$ refers to the overall baseline level across the J participants. The deviation of participant $j$ from the overall baseline level $\theta_0$ is indicated by $u_{0j}$. These deviations are assumed to be multivariate normally distributed around 0 with a variance of $\sigma_{u_0}^2$. As such, $\sigma_{u_0}^2$ represents the between-participant variance of the baseline level. Likewise, $\theta_1$ refers to the overall intervention effect across the $J$ participants; $u_{1j}$ refers to the deviation of participant $j$ from the overall intervention

[1] MVN stands for multivariate normally distributed. MVN indicates that there is covariance between the baseline and intervention effect.

effect. These deviations are assumed to be multivariate normally distributed with average of 0 and variance $\sigma_{u_1}^2$. Thus, $\sigma_{u_1}^2$ represents the between-participant variance of the intervention effect. The covariance between the baseline level and the intervention effect is indicated by $\sigma_{u_0}\, u_1$. A negative value for this covariance parameter would indicate that participants with higher baseline levels, in general, have a smaller intervention effect. The within-participant errors, $e_{ij}$, are assumed to be independent and normally distributed around 0 with a variance of $\sigma_e^2$. Therefore, $\sigma_e^2$ indicates the within-participant error variance. By substituting the Level 2 equations into the Level 1 equation, the combined two-level hierarchical linear model can be written as:

$$y_{ij} = \theta_0 + u_{0j} + \left(\theta_1 + u_{1j}\right)D_{ij} + e_{ij} \text{with} \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim MVN(0, \Sigma_v), e_{ij} \sim N\left(0, \sigma_e^2\right) \quad (3)$$

The basic two-level HLM approach introduced in Eq. 3 is promising and its appropriateness in summarizing SCED data has been empirically validated through large-scale Monte Carlo simulation studies (e.g., Ferron et al., 2009; Ferron et al., 2010; Ferron et al., 2014). The HLM approach produces unbiased estimations of the intervention effects. Ferron et al. (2009) suggests using a Kenward-Roger method of degrees of freedom to ensure the greatest accuracy of the estimation.

A major advantage of using HLM is its flexibility. For example, the basic two-level hierarchical linear model (displayed in Eq. 3) can easily be extended (Van den Noortgate & Onghena, 2003a). If a substantial amount of between-participant variability is found in intervention effectiveness (i.e., large $\sigma_{u_1}^2$), moderators can be included at the second level of the hierarchical linear model in an attempt to explain this variability. The two-level combined model, including a set of $M$ moderators at level two, is displayed in Equation 4.

$$\begin{aligned} &\mathrm{y_{ij}} = \theta_0 + \mathrm{u_{0j}} + \left(\theta_1 + \mathrm{u_{1j}}\right)\mathrm{D_{ij}} + \textstyle\sum_{\mathrm{m}=1}^{\mathrm{M}}\left(\theta_{1+\mathrm{m}} + \mathrm{u_{(1+m)j}}\right)\mathrm{M_m}\,\mathrm{D_{ij}} + \mathrm{e_{ij}}, \\ &\text{with} \begin{bmatrix} u_{0j} \\ u_{1j} \\ u_{(1+m)j} \end{bmatrix} \sim MVN(0, \Sigma_u), e_{ij} \sim N\left(0, \sigma_e^2\right) \end{aligned} \quad (4)$$

In Eq. 4, $M$ refers to the total number of moderators added to the second level ($m$ = 1 to $M$). The coefficients of the first moderator ($m$ = 1) up to the last moderator ($m$ = $M$) are $\theta_2$ , $\theta_3$, …, $\theta_{1+m}$. The coefficient $\theta_{1+m}$ indicates the overall impact of moderator m on the variability in intervention effectiveness between participants. The coefficient $u_{(1+m)j}$ refers to the difference between the overall effect of moderator m across participants and the participant-specific effect of moderator $m$ on the variability in intervention effectiveness. This study focused on explaining variability in intervention effectiveness between participants. The variability in baseline levels is assumed to be small. As such, a parameter reflecting the impact of the moderator on baseline levels is not included in Equation 4. The model presented in Eq. 4 has the potential to provide an answer to a variety of practical research questions such as: (1) Is there evidence in support of intervention effectiveness across participants? (2) For whom is the intervention most or least effective (participant-specific intervention effects)? and (3) Can variability in intervention effectiveness be explained by a set of moderators?

### Goals of Current Study

Previous methodological studies have been conducted to empirically validate the basic two-level HLM approach and several extensions to summarize single-case data (Ferron et al., 2009; Ferron et al., 2010; Ferron et al., 2014; Hembry et al., 2015). One extension that has not been investigated is the inclusion of moderators to explain variability in intervention effects between participants within a given SCED study. The current study was designed to fill this gap and has a methodological part and an applied part.

#### Methodological Part

Methodological work is needed to empirically investigate under which realistic SCED conditions (e.g., number of measurement occasions, participants, magnitude of the effects and variance) true intervention and moderator effects can be detected with sufficient power. To meet this goal, a large-scale Monte Carlo simulation study was conducted.

#### Applied Part

Although the conditions included in the simulation study are common and representative for the field, it is likely that researchers will have design conditions and parameter values deviating from the ones included in the simulation study. For this purpose, a user-friendly point-and-click Shiny tool, called *PowerSCED* (Xu et al., 2021), was developed. The tool is freely accessible, and the power is estimated based on user-defined conditions and parameter values. The first goal of the applied part was to introduce this tool. Second, a demonstration of using HLM with the inclusion of moderators was provided using data from two previously published SCED studies in the journal *School Psychology Quarterly*.

## Methodological Part

A Monte Carlo simulation study was conducted to evaluate the power of the two-level HLM approach to estimate intervention and moderator effects in conditions representative for the field of SCED research.

### Data Generation

Previous research indicates that SCED studies commonly include 0–3 moderators in an effort to explain variability in intervention effectiveness between participants (Moeyaert et al., 2021a). Moeyaert et al. (2021b) conducted a systematic review summarizing moderator characteristics in SCED research. Based on the 60 SCED meta-analyses that met their inclusion, it was found that, gender, disability type, and age are the three most frequently investigated moderators. All the identified meta-analyses were in the field of social sciences with a focus on the population of participants with special needs. The most commonly used measurement scale of the moderators gender and disability is nominal (with two categories), whereas the most commonly used measurement scale of age is continuous. Thus, four models were used

to generate MBD data. Model 0 was the basic model introduced in Eq. 2 and included no moderators. Model 1 included one binary nominal moderator (i.e., gender). Model 2 included two binary nominal moderators (i.e., gender and disability), and Model 3 contained two binary moderators (i.e., gender, disability) and one continuous moderator (i.e., age). Two binary nominal moderators were included in Model 2 and Model 3 instead of two continuous moderators. The conditions to reach a power of .80 for the model with two binary variables were more challenging compared to the model with one or two continuous variables. As such, less measurements and participants and smaller sizes of the effects would be required. Therefore, if sufficient power is reached for the suggested models, then it can be concluded that sufficient power is reached for the other models (with more continuous variables). If the applied researcher is interested in designing a study with a different combination of moderator scales than included in this simulation study, then the Shiny tool *PowerSCED* can be used.

The models are presented in Eqs. 5–8.

$$\begin{aligned} \textit{Model } 0: \quad & y_{ij} = \theta_0 + u_{0j} + \left(\theta_1 + u_{1j}\right)D_{ij} + e_{ij}, \\ & \text{with} \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N(0, \Sigma_u), e_{ij} \sim N\left(0, \sigma_e^2\right) \end{aligned} \tag{5}$$

$$\begin{aligned} \textit{Model } 1: \quad & y_{ij} = \theta_0 + u_{0j} + \left(\theta_1 + u_{1j}\right)D_{ij} + \left(\theta_2 + u_{2j}\right)\textit{Gender } D_{ij} + e_{ij}, \\ & \text{with} \begin{bmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \end{bmatrix} \sim MVN(0, \Sigma_u), \quad e_{ij} \sim N\left(0, \sigma_e^2\right) \end{aligned} \tag{6}$$

$$\begin{aligned} \textit{Model } 2: \quad & y_{ij} = \theta_0 + u_{0j} + \left(\theta_1 + u_{1j}\right)D_{ij} + \left(\theta_2 + u_{2j}\right)\textit{Gender } D_{ij} + \left(\theta_3 + u_{3j}\right)\textit{ Disability } D_{ij} + e_{ij}, \\ & \text{with} \begin{bmatrix} u_{0j} \\ u_{1j} \\ u_{3j} \\ u_{3j} \end{bmatrix} \sim MVN(0, \Sigma_u), e_{ij} \sim N\left(0, \sigma_e^2\right) \end{aligned} \tag{7}$$

$$\begin{aligned} \textit{Model } 3: & y_{ij} = \theta_0 + u_{0j} + \left(\theta_1 + u_{1j}\right)D_{ij} + \left(\theta_2 + u_{2j}\right)\textit{Gender } D_{ij} + \left(\theta_3 + u_{3j}\right)\textit{ Disability } D_{ij} + \left(\theta_4 + u_{4j}\right)\textit{ Age} D_{ij} + e_{ij}, \\ & \text{with} \begin{bmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \\ u_{3j} \\ u_{4j} \end{bmatrix} \sim MVN(0, \Sigma_u), e_{ij} \sim N\left(0, \sigma_e^2\right) \end{aligned} \tag{8}$$

The Level-2 residuals ($u$'s) in Eqs. 5–8 were assumed to be multivariate normally distributed. The Level-1 residuals ($e_{ij}$'s) in Eqs. 5–8 were assumed to be independent and normally distributed. This means that it was assumed that outcome scores closer in time were not more related to each other compared to outcomes further away in time. This assumption can be violated in repeated measured design such as SCEDs, and

dependency between errors (i.e., autocorrelation) can be modeled (Ferron et al., 2009; McKnight et al., 2000; Petit-Bois et al., 2016). However, autocorrelation was not included in current study because of the following reasons: First, Shadish and Sullivan (2011) conducted a systematic review of SCEDs, and summarized SCED data characteristics. One of their findings indicated that the size of autocorrelation in SCED studies varies tremendously, with an average of about zero. For that reason, assuming an autocorrelation of 0 is a reasonable assumption to start with. Second, methodological research focusing on modeling moderators in SCEDs is limited. Therefore, this modeling complexity needs to be studied in isolation to avoid confounding with other modeling complexities. Finally, there are many covariance structures possible, of which the first-order autoregressive type is most common (Ferron et al., 2009; McKnight et al., 2000; Petit-Bois et al., 2016). Given that there are multiple functional forms possible and plausible values for the autocorrelation parameter, more research is needed to first focus on this complexity in isolation. Therefore, the issue of autocorrelation deserves its own study. This last point is further motivated by Petit-Bois et al. (2016) explicitly stating: "research will be needed to fully understand the effects of various types of model misspecification on the fixed-effect inferences made from single-case data" (p. 810).

The following factors were varied to simulate the SCED data: (1) the true values of the intervention effect $\theta_1$ and moderator effects $\theta_{1+m}$; (2) the number of Level-1 and Level-2 units (i.e., the number of measurement occasions, $I$, and the number of participants, $J$, respectively); (3) the number of moderators at the second level; and (4) the covariance between moderators and the covariance between moderator and intervention.

To obtain realistic values of moderator effects, variability in moderator effects, covariances between moderators, and covariances between moderator and intervention, primary SCED studies from two published SCED meta-analyses were reanalyzed. The two SCED meta-analyses selected for this purpose were Moeyaert et al. (2021b) and Heyvaert et al. (2012).[2] Standardized SCED data from 45 SCED studies and 349 participants from Moeyaert et al. (2021b), and 216 SCED studies and 469 participants from Heyvaert et al. (2012) were reanalyzed to identify realistic parameter values for moderator effects, variability in moderator effects, and covariances between moderators and covariances between moderator and intervention. Covariances between moderators and between moderator and intervention were set to either all zero or nonzero (i.e., realistic values; see Table 1). Values for the other design conditions (intervention effect, the number of measurement occasions and number of participants) were based on previous methodological work in the field of HLM of SCEDs (Ferron et al., 2014; Moeyaert et al., 2013a, 2013b, see Table 1). The amount of between-participant variance of the baseline, the intervention effect and the moderator effects, and the within-participant variance were kept fixed (see Table 1). This resulted in 16 conditions for Model 0, 64 conditions for Model 1, 128 conditions for Model 2, and 256 conditions for Model 3. An overview of the values per varying design factors is given in Table 1. An equal number of participants were assigned to the categories of the two dichotomous moderator variables (i.e., gender and disability type). This was to avoid

[2] SCED data were standardized prior to re-analysis (Van den Noortgate & Onghena, 2008). This took into account that primary SCED studies might use different outcome scales. For instance, one SCED study might measure the outcome on a scale from 0 to 10 whereas another SCED study might use a scale from 0 to 100.

that all participants within one study fell within one category and as such modeling a moderator would not make sense. The same was true for the continuous variable age (i.e., not all participants within one study had the same age). Values for the moderator variable age were generated with values ranging from 5 to 15 (based on reanalyzing 349 participants included in the meta-analysis by Moeyaert et al., 2021b), and the variable was mean-centered in order to provide a meaningful interpretation of the intercept. One thousand replications were generated for each condition, leading to a total of 464,000 datasets to analyze. The annotated SAS code used to generate the SCED data can be obtained by contacting the first author of this study.

The data was generated for MBD across participants; thus, the start point of the intervention was staggered and was dependent on the number of participants and measurements (e.g., the intervention started at 7, 10, 13 and 16, respectively for studies with four participants and 20 measurement occasions).

## Data Analysis

The same four two-level hierarchical linear models used to generate the SCED data (see Eqs. 5–8) were used to estimate the intervention and moderator effects. For this purpose, the PROC MIXED procedure within SAS (SAS Institute Inc., 2013) was used. The restricted maximum likelihood (REML) via the Kenward-Rogers method for

**Table 1** Overview of Design Factors and Parameter Values

| Design Factor | Notation | Value |
|---|---|---|
| Number of measurement occasions | $I$ | 20 or 40 |
| Number of participants | $J$ | 4, 7, 12, or 20 |
| Intervention effect | $\theta_1$ | 3 or 4 |
| Moderator effects | Gender, $\theta_2$ | 1 or 3 |
| | Disability, $\theta_3$ | 0.75 or 1.50 |
| | Age, $\theta_4$ | 0.25 or 0.50 |
| Covariance | Gender and disability, $\sigma^2_{GD}$ | 0 or 0.20 |
| | Gender and age, $\sigma^2_{GA}$ | 0 or 0.02 |
| | Disability and age, $\sigma^2_{DA}$ | 0 or 0.02 |
| | Intervention and gender, $\sigma^2_{IG}$ | 0 or 0.05 |
| | Intervention and disability, $\sigma^2_{ID}$ | 0 or 0.01 |
| | Intervention and age, $\sigma^2_{IA}$ | 0 or 0.01 |
| Between-participant variance | Baseline level, $\sigma^2_{\theta 0}$ | 1 |
| | Intervention effect, $\sigma^2_{\theta 1}$ | 1 |
| | Gender moderator effect, $\sigma^2_{\theta 2}$ | 1 |
| | Disability moderator effect, $\sigma^2_{\theta 3}$ | 0.5 |
| | Age moderator effect, $\sigma^2_{\theta 4}$ | 0.2 |
| Within-participant variance | $\sigma^2_e$ | 1 |

*Note*. $\sigma^2_{GD}$ = Covariance between gender and disability; $\sigma^2_{GA}$ = Covariance between gender and age; $\sigma^2_{DA}$ = Covariance between disability and age; $\sigma^2_{IG}$ = Covariance between intervention and gender; $\sigma^2_{IG}$ = Covariance between intervention and disability; $\theta^2$ = Covariance between intervention and age

degrees of freedom was chosen, because this is the recommended estimation procedure in contexts of a small number of participants and measurement occasions within participants (Ferron et al., 2009). The aim of the simulation study was to investigate the power to estimate intervention and moderator effects under commonly encountered SCED study conditions. The power is the probability of detecting a statistically significant intervention or moderation effects for nonzero true effects. Thus, power was calculated for the design conditions having nonzero effects. In particular, when the null hypothesis (there was no true intervention or moderation effects) was false, and a statistically significant effect was found ($p < .05$), then the nonzero effect had been successfully detected. If the null hypothesis was not rejected (assuming the null hypothesis is false), then a Type II error was made. The percentage of replications per condition in which the nonzero significant intervention or moderation effects were successfully detected (i.e., $\frac{\#successfully\ rejected\ replications}{1{,}000}$) was the power to detect the intervention or moderator effects. The performance of the HLM approach with moderators was further investigated by examining the following statistical properties of the intervention effect and moderator effects: relative bias, mean squared error, relative standard error bias, coverage proportion of the 95% confidence interval, and Type I error rate. These results can be obtained by contacting the first author of this study. Generalized linear modeling (GLM) was used to identify design factors that have a statistically significant and large impact on the power. As such, power was modeled as dependent variable and the varying design factors as independent variables (both main and two-way interaction effects were modeled). This helped investigating the influence of design factors on power, and identify conditions in which the two-level HLM approach had sufficient power (power larger than or equal to .80) to estimate intervention and moderator effects. A large impact was indicated by a partial eta-squared ($\eta_p^2$) value larger than or equal to .14 (Cohen, 1969).

## Results

This section presented the power to estimate intervention and moderator effects per model. As mentioned before, Models 0–3 included no moderator, one binary moderator, two binary moderators, and two binary and one continuous moderator, respectively.

When there were four study participants, Models 2 and 3 (including more than one categorical moderator) failed to estimate the second positioned categorical moderator. However, the continuous moderator (i.e., age in this study) could still be successfully estimated regardless of its position in the model. For instance, when Model 2 included four participants, and gender took the first position and disability the second position, Model 2 only provided an estimate of the effect of gender. Likewise, when Model 3 included only four participants, and the three moderators were modeled in the following sequence: age, gender, and disability type, Model 3 failed to estimate disability type. If the moderators were in the sequence gender, disability type, and age, Model 3 still failed to estimate disability type, but not age. The possible reason was that the categorical moderators only had two levels (0 or 1), and as such the true moderator effects were smaller compared to the true moderator effect of the continuous moderator (i.e., age having a range of 5–15). Therefore, the conditions with four participants in

Models 2 and 3 were not included. As a consequence, a total 16 design conditions for Model 0, 64 design conditions for Model 1, 96 design conditions for Model 2, and 192 design conditions for Model 3 were analyzed and discussed. The results were not presented per level of the covariance factor because it had no main effect (and did not interact with other factors) on the power estimates. The interested reader can obtain these results by contacting the first author.

### Model 0: No Moderators

The GLM results indicate that all design factors had a large effect on the power to estimate the intervention effect ($.617 < \eta_p^2 < .999$). The power to detect the intervention effect per design condition is presented in Table 2. All conditions had a power larger than .80. For four participants, the power increased from .94 to 1.00, when the intervention effect $\theta_1$ increased from 3 to 4. The larger the magnitude of the intervention effect, the larger the power. When the number of participants was 7, 12, or 20, no differences in power was obtained between $\theta_1 = 3$ and $\theta_1 = 4$ as the power reached 1.00 in all conditions.

### Model 1: One Categorical Moderator

The GLM results indicate that the number of participants (for intervention effect, $\eta_p^2 = .998$; for gender moderator effect, $\eta_p^2 = .997$), the number of measurement occasions (for intervention effect, $\eta_p^2 = .481$; for gender moderator effect, $\eta_p^2 = .309$), and the magnitude of intervention effect ($\eta_p^2 = .976$) and gender moderator effect ($\eta_p^2 = .998$) had large effects on the power to detect the intervention effect and gender moderator effect. More participants, more measurement occasions, and a larger magnitude of the intervention effect or gender moderator effect led towards a larger power.

The power to detect the nonzero intervention and moderation effect across the 32 design conditions (controlling for the factor covariance) is presented in Table 3. The results indicate that, across all conditions, the power to detect the intervention effect was larger compared to the power to detect the moderator effect. Studies with seven participants or more had a power larger than .80 to detect the intervention effect across

**Table 2** Power Estimates to Detect Intervention Effect—No Moderators

| *I* | *J* | $\theta_1 = 3$ | $\theta_1 = 4$ |
|---|---|---|---|
| 20 | 4 | .94 | 1.00 |
| | 7 | 1.00 | 1.00 |
| | 12 | 1.00 | 1.00 |
| | 20 | 1.00 | 1.00 |
| 40 | 4 | .94 | 1.00 |
| | 7 | 1.00 | 1.00 |
| | 12 | 1.00 | 1.00 |
| | 20 | 1.00 | 1.00 |

*Note*. $\theta_1$ = Intervention effect; *I* = Number of measurement occasions; *J* = Number of participants

all conditions. For the moderator effect, only studies with a relatively large true moderation effect (i.e., $\theta_2$ = 3) and 12 or more participants had sufficient power to detect the moderator effect.

### Model 2: Two Categorical Moderators

Similar to Model 1, the number of participants (.996 < $\eta_p^2$ < 1.00), the number of measurement occasions (.436 < $\eta_p^2$ < .638), and the magnitude of intervention effect ($\eta_p^2$ = .976), and moderator effects ($\eta_p^2$ = .999 for gender and $\eta_p^2$ = .998 for disability) largely affected the power to detect corresponding intervention and moderator effects. More participants, more measurement occasions, and larger sizes of the intervention effect and moderator effects resulted in larger power.

The power to detect the nonzero intervention effect, and the two moderator effects per design condition is presented in Table 2. Similar to Model 1, the power to detect the intervention effect was larger compared to the power to estimate the moderator effects. According to Table 4, studies with 12 or more participants had a sufficient power to detect the intervention effect (regardless of the true value of the intervention effect is 3 or 4). However, to detect the first moderator effect (i.e., gender), studies needed at least 12 participants and a relatively large true moderator effect (i.e., $\theta_2$ = 3) to reach a power of .80. To detect the second moderator (i.e., disability) with sufficient power, studies required a large true moderator effect (i.e., $\theta_3$ = 1.50) and at least 20 participants.

### Model 3: Two Categorical and One Continuous Moderator

The GLM results indicate that the number of participants (.99 < $\eta_p^2$ < 1.00), the number of measurement occasions (.25 < $\eta_p^2$ < .46), and the magnitude of intervention effect and moderator effects (.95 < $\eta_p^2$ < 1.00) largely affected the power to estimate

**Table 3** Power Estimates to Estimate Intervention and Moderator Effect—One Moderator

| Power_ D | | | | | | | Power_ Gender | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *I* | θ1 | θ2 | *J* = 4 | *J* = 7 | *J* = 12 | *J* = 20 | *J* = 4 | *J* = 7 | *J* = 12 | *J* = 20 |
| 20 | 3 | 1 | .55 | **.87** | **1.00** | **1.00** | .11 | .13 | .29 | .46 |
| | | 3 | .54 | **.85** | **1.00** | **1.00** | .33 | .73 | **.97** | **1.00** |
| | 4 | 1 | .74 | **.98** | **1.00** | **1.00** | .11 | .15 | .26 | .46 |
| | | 3 | .73 | **.98** | **1.00** | **1.00** | .32 | .76 | **.98** | **1.00** |
| 40 | 3 | 1 | .56 | **.88** | **1.00** | **1.00** | .10 | .15 | .28 | .49 |
| | | 3 | .56 | **.90** | **1.00** | **1.00** | .35 | .77 | **.98** | **1.00** |
| | 4 | 1 | .78 | **.99** | **1.00** | **1.00** | .12 | .15 | .28 | .49 |
| | | 3 | .79 | **.99** | **1.00** | **1.00** | .35 | .77 | **.99** | **1.00** |

*Note 1*. Power_D = Power to estimate intervention; Power_Gender = Power to estimate moderator (gender)
*Note 2*. $\theta_1$ = Intervention effect; $\theta_2$ = Moderation effect (gender); *I* = Number of measurement occasions; *J* = Number of participants

*Note 3*. Values larger than or equal to .80 are indicated in bold

corresponding intervention effect and moderator effects. Consistent with Model 0, Model 1, and Model 2, all four models revealed a similar pattern in that studies with a larger number of participants, measurement occasions, and a larger value of the magnitude of intervention effect and moderator effects, had a larger power to detect corresponding intervention and moderator effects.

The power to detect the true intervention and moderator effects across 96 design conditions (controlling for the factor covariance) is presented in Table 3. As displayed in Table 5, studies with 12 or more participants had a power of .80 or larger to detect the intervention effect across different magnitudes of the intervention effect and the numbers of measurement occasions. For the first categorical moderator (i.e., gender), studies with 12 or more participants and a true moderator effect of 3 reached a power of .80 regardless the number of measurement occasions. Likewise, the continuous moderator (i.e., age) would be detected with sufficient power for studies with at least 12 participants and an effect of 0.50. Whereas, for the second categorical moderator (i.e., disability), studies with 20 participants, 40 measurement occasions, and an effect of 1.50 were needed to obtain sufficient power to capture the second moderator effect. Overall, for a study to have sufficient power to estimate the intervention effect and all three moderators effects, 20 participants, 40 measurement occasions, and relatively large sizes of all moderator effects (i.e., $\theta_2 = 3$, $\theta_3 = 1.50$, $\theta_4 = 0.50$) were needed.

**Table 4** Power to Estimate the Intervention and Moderator Effects—Two Moderators

| | | | | Power_D | | | Power_Gender | | | Power_Disability | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *I* | $\theta_1$ | $\theta_2$ | $\theta_3$ | *J* = 7 | *J* = 12 | *J* = 20 | *J* = 7 | *J* = 12 | *J* = 20 | *J* = 7 | *J* = 12 | *J* = 20 |
| 20 | 3 | 1 | 0.75 | .44 | **.99** | **1.00** | .13 | .24 | .45 | .09 | .17 | .29 |
| | | | 1.50 | .45 | **.99** | **1.00** | .11 | .26 | .47 | .21 | .55 | **.80** |
| | | 3 | 0.75 | .44 | **.98** | **1.00** | .52 | **.98** | **1.00** | .10 | .16 | .30 |
| | | | 1.50 | .43 | **.98** | **1.00** | .52 | **.97** | **1.00** | .21 | .54 | **.80** |
| | 4 | 1 | 0.75 | .66 | **1.00** | **1.00** | .13 | .24 | .47 | .09 | .18 | .30 |
| | | | 1.50 | .68 | **1.00** | **1.00** | .12 | .25 | .45 | .20 | .53 | **.80** |
| | | 3 | 0.75 | .64 | **1.00** | **1.00** | .53 | **.98** | **1.00** | .09 | .18 | .30 |
| | | | 1.50 | .65 | **1.00** | **1.00** | .53 | **.97** | **1.00** | .22 | .53 | **.82** |
| 40 | 3 | 1 | 0.75 | .49 | **.99** | **1.00** | .13 | .26 | .51 | .09 | .19 | .31 |
| | | | 1.50 | .50 | **.99** | **1.00** | .11 | .27 | .50 | .24 | .57 | **.84** |
| | | 3 | 0.75 | .47 | **.99** | **1.00** | .56 | **.99** | **1.00** | .10 | .19 | .31 |
| | | | 1.50 | .47 | **1.00** | **1.00** | .56 | **.98** | **1.00** | .23 | .57 | **.84** |
| | 4 | 1 | 0.75 | .71 | **1.00** | **1.00** | .12 | .29 | .51 | .09 | .19 | .32 |
| | | | 1.50 | .70 | **1.00** | **1.00** | .12 | .28 | .49 | .23 | .59 | **.82** |
| | | 3 | 0.75 | .70 | **1.00** | **1.00** | .59 | **.98** | **1.00** | .10 | .19 | .31 |
| | | | 1.50 | .70 | **1.00** | **1.00** | .55 | **.99** | **1.00** | .22 | .60 | **.84** |

*Note 1*. Power_D = Power to estimate intervention; Power_Gender = Power to estimate moderator 1 (gender); Power_Disability = Power to estimate moderator 2 (disability)

*Note 2*. $\theta_1$ = Intervention effect; $\theta_2$ = Moderator 1 effect (gender); $\theta_3$ = Moderator 2 effect (disability); *I* = Number of measurement occasions; *J* = Number of participants

*Note 3*. Values larger than or equal to .80 are indicated in bold

**Table 5** Power to Estimate Intervention and Moderator Effects—Three Moderators

| | | | | | Power_D | | | Power_Gender | | | Power_Disability | | | Power_Age | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *I* | θ1 | θ2 | θ3 | θ4 | *J* = 7 | *J* = 12 | *J* = 20 | *J* = 7 | *J* = 12 | *J* = 20 | *J* = 7 | *J* = 12 | *J* = 20 | *J* = 7 | *J* = 12 | *J* = 20 |
| 20 | 3 | 1 | 0.75 | 0.25 | .28 | **.93** | **1.00** | .09 | .23 | .46 | .09 | .18 | .26 | .16 | .42 | .70 |
| | | | | 0.50 | .30 | **.93** | **1.00** | .09 | .23 | .46 | .08 | .16 | .25 | .42 | **.88** | **.99** |
| | | | 1.50 | 0.25 | .30 | **.92** | **1.00** | .09 | .24 | .44 | .14 | .50 | .79 | .16 | .46 | .68 |
| | | | | 0.50 | .31 | **.93** | **1.00** | .11 | .22 | .43 | .18 | .48 | .76 | .43 | **.89** | **.99** |
| | | 3 | 0.75 | 0.25 | .29 | **.95** | **1.00** | .37 | **.95** | **1.00** | .08 | .16 | .25 | .16 | .41 | .71 |
| | | | | 0.50 | .33 | **.93** | **1.00** | .37 | **.95** | **1.00** | .08 | .17 | .24 | .43 | **.90** | **1.00** |
| | | | 1.50 | 0.25 | .33 | **.92** | **1.00** | .36 | **.95** | **1.00** | .14 | .48 | .77 | .17 | .38 | .69 |
| | | | | 0.50 | .29 | **.94** | **1.00** | .38 | **.94** | **1.00** | .16 | .48 | .79 | .41 | **.89** | **.99** |
| | 4 | 1 | 0.75 | 0.25 | .45 | **.99** | **1.00** | .10 | .26 | .45 | .09 | .14 | .27 | .17 | .42 | .67 |
| | | | | 0.50 | .46 | **.99** | **1.00** | .11 | .24 | .44 | .09 | .17 | .27 | .42 | **.90** | **1.00** |
| | | | 1.50 | 0.25 | .46 | **.99** | **1.00** | .10 | .25 | .46 | .17 | .47 | .77 | .15 | .41 | .71 |
| | | | | 0.50 | .47 | **.98** | **1.00** | .10 | .23 | .45 | .16 | .49 | .78 | .42 | **.88** | **.99** |
| | | 3 | 0.75 | 0.25 | .48 | **.99** | **1.00** | .39 | **.93** | **1.00** | .08 | .14 | .29 | .17 | .39 | .70 |
| | | | | 0.50 | .46 | **.99** | **1.00** | .38 | **.95** | **1.00** | .09 | .15 | .25 | .38 | **.88** | **1.00** |
| | | | 1.50 | 0.25 | .45 | **.99** | **1.00** | .41 | **.95** | **1.00** | .15 | .48 | .79 | .15 | .39 | .69 |
| | | | | 0.50 | .46 | **.99** | **1.00** | .38 | **.94** | **1.00** | .15 | .46 | .78 | .41 | **.89** | **.99** |
| 40 | 3 | 1 | 0.75 | 0.25 | .30 | **.96** | **1.00** | .10 | .25 | .47 | .09 | .20 | .28 | .17 | .45 | .74 |
| | | | | 0.50 | .32 | **.95** | **1.00** | .08 | .24 | .47 | .08 | .18 | .27 | .46 | **.90** | **1.00** |
| | | | 1.50 | 0.25 | .32 | **.94** | **1.00** | .10 | .25 | .48 | .16 | .53 | **.83** | .18 | .47 | .73 |
| | | | | 0.50 | .32 | **.94** | **1.00** | .10 | .25 | .47 | .19 | .51 | **.80** | .45 | **.91** | **.99** |
| | | 3 | 0.75 | 0.25 | .31 | **.96** | **1.00** | .40 | **.96** | **1.00** | .08 | .16 | .30 | .16 | .45 | .75 |
| | | | | 0.50 | .33 | **.96** | **1.00** | .40 | **.96** | **1.00** | .08 | .16 | .26 | .44 | **.92** | **1.00** |
| | | | 1.50 | 0.25 | .31 | **.94** | **1.00** | .38 | **.96** | **1.00** | .15 | .52 | **.81** | .16 | .43 | .75 |
| | | | | 0.50 | .31 | **.95** | **1.00** | .38 | **.96** | **1.00** | .17 | .53 | **.80** | .44 | **.90** | **1.00** |
| | 4 | 1 | 0.75 | 0.25 | .50 | **.99** | **1.00** | .10 | .27 | .49 | .09 | .18 | .27 | .17 | .46 | .73 |
| | | | | 0.50 | .48 | **.99** | **1.00** | .11 | .25 | .48 | .08 | .18 | .28 | .44 | **.91** | **1.00** |
| | | | 1.50 | 0.25 | .45 | **.99** | **1.00** | .10 | .24 | .48 | .17 | .50 | **.82** | .16 | .46 | .74 |
| | | | | 0.50 | .46 | **.99** | **1.00** | .10 | .26 | .45 | .16 | .53 | **.81** | .45 | **.90** | **.99** |
| | | 3 | 0.75 | 0.25 | .51 | **1.00** | **1.00** | .43 | **.95** | **1.00** | .09 | .15 | .30 | .19 | .44 | .72 |
| | | | | 0.50 | .48 | **.99** | **1.00** | .43 | **.97** | **1.00** | .09 | .15 | .28 | .45 | **.91** | **1.00** |
| | | | 1.50 | 0.25 | .48 | **1.00** | **1.00** | .41 | **.96** | **1.00** | .15 | .52 | **.82** | .17 | .43 | .74 |
| | | | | 0.50 | .48 | **.99** | **1.00** | .40 | **.97** | **1.00** | .16 | .50 | **.82** | .44 | **.91** | **1.00** |

*Note 1*. Power_D = Power to estimate intervention; Power_Gender = Power to estimate Moderator 1 (gender); Power_Disability = Power to estimate Moderator 2 (disability); Power_Age = Power to estimate Moderator 3 (age)

*Note 2*. $\theta_1$ = Intervention effect; $\theta_2$ = Moderator 1 effect (gender); $\theta_3$ = Moderator 2 effect (disability); $\theta_4$ = Moderator 3 effect (age); *I* = Number of measurement occasions; *J* = Number of participants

*Note 3*. Values larger than or nearly equal to .80 are indicated in bold

## Applied Part

### PowerSCED Shiny Tool

*PowerSCED* is a Shiny application developed to estimate the power to detecting a nonzero intervention effect and nonzero moderator effects in MBD across participant's studies. The *PowerSCED* tool was developed for the two-level HLM approach, with the option to quantify the intervention effect as a change in level and/or a change in slope. This tool can assist applied researchers in designing a MBD study with sufficient power to detect true intervention and moderator effects using two level HLM. The steps to use *PowerSCED* for the two-level HLM with the inclusion of moderators are presented in Fig. 2. First, the user defines the specific model (i.e., change in level and/or change in trends). Next, the user defines the design condition (i.e., number of participants, number of measurement occasions per condition, and the number of moderators). Finally, the user includes anticipated values for the intervention effect, moderator effect(s), and the between and within (co)variances. Once the model is defined, the design condition is defined, and the parameter values are set, the *PowerSCED* application provides power estimates for the intervention effect and the moderator effect(s) (using the Monte Carlo simulation technique based on 1,000 replications). Researchers are recommended to use *PowerSCED* before conducting their study to ensure sufficient power. Researchers can also use the *PowerSCED* tool post hoc to explore whether the nonsignificant effect(s) is(are) due to insufficient power. An empirical demonstration of the two-level HLM approach with the inclusion of moderator(s) is presented in the next section. A post-hoc power analysis was conducted as the two empirical studies have been completed.

### Demonstration of the Use of Two-Level HLM with the Inclusion of Moderators

Two empirical SCED studies published in *School Psychology Quarterly* were selected for the empirical demonstration of the two-level HLM approach, with the inclusion of moderators, to summarize SCED effects. The first study by King et al. (2017) included six participants (five males and one female; five Caucasian and one Hispanic) to examine the effectiveness of the On-Task in a Box program for increasing on-task behavior of highly off-task students. Note that both moderators were highly imbalanced and not representing a condition embedded in the simulation study. The influence of imbalanced moderators on the power is beyond the scope of this article, and a direction for future research. The second study by Shernoff and Kratochwill (2007) included 13 participants (eight males and five females; nine European American and four others, such as African American and Asian). Shernoff and Kratochwill examined the effectiveness of self-administered videotape modeling for reducing preschoolers' disruptive behavior problems. These two SCED studies both used the MBD across participants to investigate the effectiveness of the intervention of interested.

The data of the two selected SCED studies were standardized before running the analyses (Van den Noortgate & Onghena, 2008). This ensured that the size of the intervention and moderator effects were comparable (i.e., at the same scale) to the ones included in the simulation study. Two-level HLM was then used to estimate the standardized intervention and moderators' effects. The results are summarized and

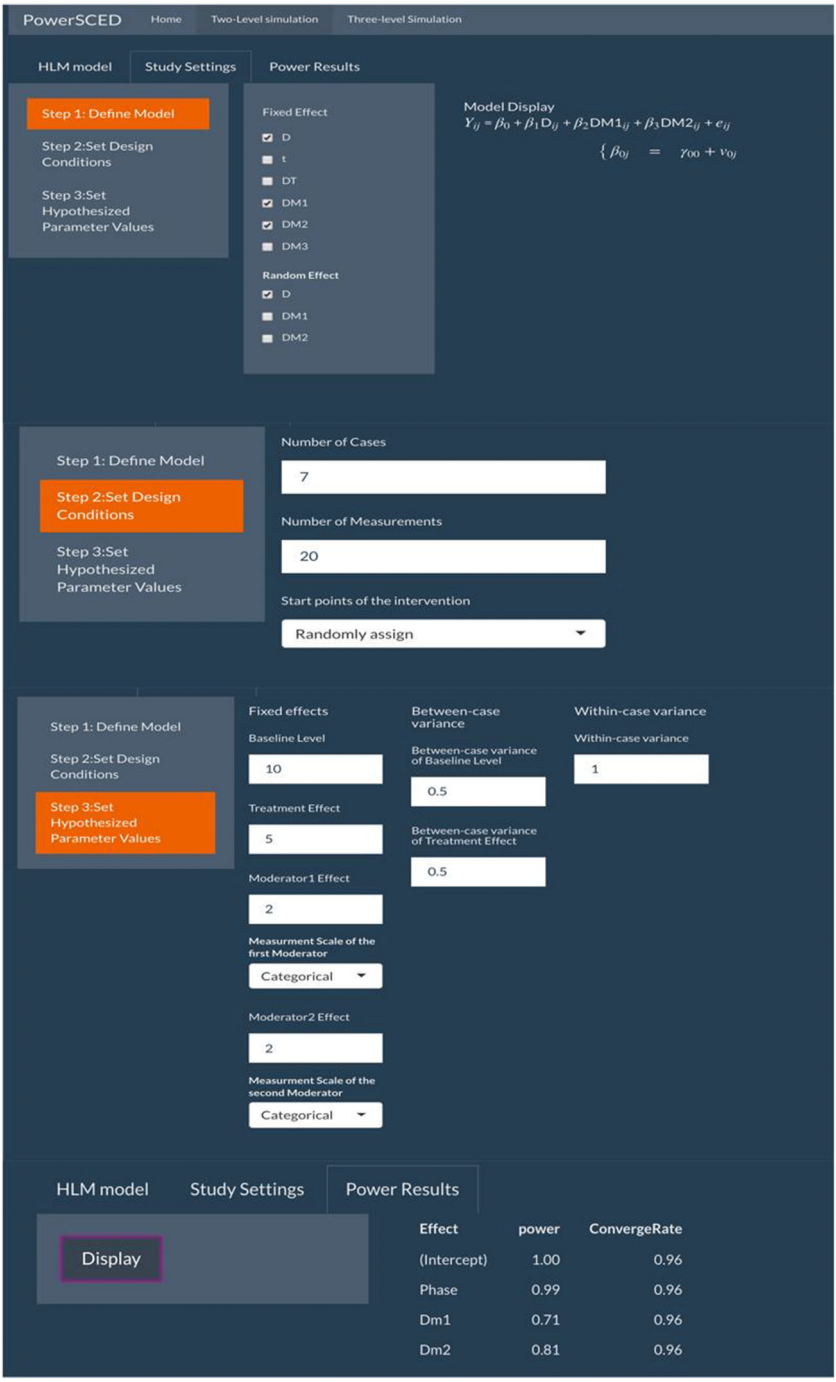


**Fig. 2** The Steps of Using PowerSCED. *Note*. *D* = Intervention effect; *DT* = Intervention trend effect; *DM1* = Moderator 1 effect; *DM2* = Moderator 2 effect; *DM3* = Moderator 3 effect

presented in Table 6. The annotated SAS code for running the two-level HLM with the inclusion of two moderators (i.e., gender and race) together with the two data sets can be obtained by contacting the first author of this article. This allows the interested reader to replicate the HLM analysis. The obtained parameter estimates were further used as hypothesized parameter values for the post-hoc power analysis using the *PowerSCED* tool.

As shown in Table 6, the intervention effect (defined as a change in level) was not statistically significant in King et al. (2017) (standardized effect size = 3.52, *SE* = 1.97, $p = .17$), whereas the intervention effect was statistically significant in Shernoff and Kratochwill (2007) (standardized effect size = -1.46, *SE* = 0.27, $p < .001$). This indicates that the On-Task in a Box program used in King et al. (2017) did not significantly increase students' on-task behavior. Whereas for the study by Schernoff and Kratochwill, the self-administered videotape modeling significantly reduced students' disruptive behavior problems. The effects of the moderators gender and race were not statistically significant in both studies. In particular, in King et al. (2017), the effect of gender was -2.21 (*SE* = 4.35, $p = .65$). Gender was a dummy coded variable with 0 indicating male participants and 1 indicating female participants. Controlling for race, the intervention increased on-task behavior for female participants with 1.31 (i.e., 3.52-2.21), whereas the on-task behavior was increased with 3.52 standardized units for male participants. Although the difference in intervention effectiveness between male and female was not statistically significant, a standardized difference of 2.21 can be considered as being practical significance. This means that for participants having the same race, the On-Task in a Box program might be more effective to increase participant's on-task behavior compared to the female participant's. Note that the *p*-value for two tailed testing the gender effect was .65, and this was an artifact of the large standard error. The effect of race in King et al. (2017) is -2.76 (*SE* = 4.35, $p =$

**Table 6** Parameter and Standard Error Estimates Resulting from Two-Level HLM Using King et al. (2017) and Shernoff and Kratochwill (2007) Data

| | | Parameter | Parameter estimate | SE | p |
|---|---|---|---|---|---|
| King et al. (2017) | Average baseline level | $\theta_0$ | 2.09 | 0.29 | < .0001 |
| | Average intervention effect | $\theta_1$ | 3.52 | 1.97 | .168 |
| | Average gender moderator effect | $\theta_2$ | -2.21 | 4.35 | .645 |
| | Average race moderator effect | $\theta_3$ | -2.76 | 4.35 | .571 |
| | Between-participants variance of intervention effect | $\sigma^2_{\theta 1}$ | 14.94 | | |
| | Within-participants variance | $\sigma^2_e$ | 2.01 | | |
| Shernoff and Kratochwill (2007) | Average baseline level | $\theta_0$ | 2.97 | 0.16 | < .0001 |
| | Average intervention effect | $\theta 1$ | -1.46 | 0.27 | < .0001 |
| | Average gender moderator effect | $\theta_2$ | -0.24 | 0.35 | .501 |
| | Average race moderator effect | $\theta_3$ | 0.55 | 0.37 | .169 |
| | Between-participants variance of intervention effect | $\sigma^2_{\theta 1}$ | 0.16 | | |
| | Within-participants variance | $\sigma^2_e$ | 1.21 | | |

.57), indicating that after controlling for gender, the intervention increased on-task behavior for one Hispanic participant with 0.76 standardized unit (i.e., 3.52-2.76), whereas for Caucasian participants the on-task behavior was increased with 3.52 standardized units A standardized difference in intervention effectiveness of 2.76 between one Hispanic participant and other Caucasian participants was obtained, though this difference was not statistically significant. This means that for participants having the same gender, the intervention might be more effective to increase Caucasian participant's on-task behavior compared to Hispanic participant's.

The effect of gender and the effect of race in Shernoff and Kratochwill (2007) were -0.24 ($SE$ = 0.35, $p$ = .50) and 0.55 ($SE$ = 0.37, $p$ = .17), respectively. Controlling for race, the intervention reduced female participants' disruptive behavior problems with 1.70, whereas male participants' disruptive behavior problems were reduced with 1.46 standardized units. A standardized difference in intervention effectiveness of 0.24 between female and male participants was obtained which was small and not statistically significant. That being said, when participants have the same race, the self-administered videotape modeling intervention might have a stronger effect on reducing male participant's disruptive behavior problems compared to female participant's. Controlling for gender, the intervention reduced disruptive behavior problems for non-European American participants (such as African American and Asian participants) with 2.01, whereas European American participants' disruptive behavior problems were reduced with 1.46. A standardized difference in intervention effectiveness of 0.55 between European American participants and other ethnical participants was small and not statistically significant. This reflects that the intervention might have a stronger effect on reducing non-European American participants' disruptive behavior problems than European American participants', when the participants had the same gender. The estimated parameter values of the between-participants variance for intervention effect and the within-participants variance are also presented in Table 6. For a more in depth discussion of the interpretation of HLM parameter estimated in the context of SCEDs we refer the interested reader to Declercq et al. (2020).

The post-hoc power analysis (using *PowerSCED)* was conducted to evaluate whether the two studies were designed in a way that they were able to capture the true moderator effects. The following design conditions were specified: $I$ = 15 and $J$ = 6 for King et al. (2017) and $I$ = 12 and $J$ = 13 for Shernoff and Kratochwill (2007). The parameters values (i.e., baseline level, intervention effect, gender and race moderator effects, between-participants variance of the intervention, and within-participants variance) were set using the values presented in Table 6. The between-participants variance of the baseline and moderator effects were set to zero as this was negligible.

The *PowerSCED* tool was used to estimate the power in detecting intervention and moderator effects. The aforementioned steps were executed to set the model, design conditions and hypothesized parameter values. The estimated power estimates are presented in Table 7.

As shown in Table 7, King et al. (2017) had a power of .34 to detect the intervention effect with a standardized magnitude of 3.52 and power of .19 and .22 to detect gender and race moderator effects, respectively. Therefore it is possible that the nonsignificant values for intervention effect and moderator effects were due to lack of power. The magnitude of the hypothesized gender and race moderator effects were around -2.21 and -2.76. All hypothesized effects were moderate to large relative to the values used in

this Monte Carlo simulation study. The reason for the low power is likely the small number of participants ($J = 6$) and large between-participants variance of the intervention ($\sigma^2_{\theta 1} = 14.94$). The post-hoc power analysis was replicated for a larger number of participants (i.e., $J = 20$) and a smaller between-participant variance of the intervention (i.e., $\sigma^2_{\theta 1} = 2$), keeping all other factors constant. For this condition, the power to estimate the intervention effect increased to 1.00 and the power to estimate gender and race moderator effects also reached a power larger than .80 (.88, and .97, respectively).

Table 7 also shows that Shernoff and Kratochwill (2007) had a power of .98 to detect an intervention effect with a standardized magnitude of -1.46. It means that for a true intervention effect with standardized size of -1.46, 98% of the studies like Schernoff and Kratochwill would be able to successfully detect the intervention effect. However, the power in the study by Shernoff and Kratochwill to detect gender and race moderator effects were low (i.e., .14 and .39, respectively). The reason for the low power is the extremely small moderator effects (for gender: -0.24; for race: 0.55). If the true value of the moderator effect of gender and race would be increased to 1.5, a power of .98 would be obtained to capture the moderator effects while keeping other factors the same.

Taken together, the two empirical studies have shown that the power to detect intervention and moderator effects were highly depended on the number of participants, the between-participant variance of intervention, and/or the magnitude of the corresponding moderator effects. Small number of participants (e.g., $J = 6$), large between-participant variance of intervention (e.g., $\sigma^2_{\theta 1} = 14.94$), and small moderator effects (e.g., -0.24 and 0.55) may result in insufficient power to capture the true intervention and moderator effects. When the true moderator effects are relatively large, the research design factors such as the number of participants can largely affect the power.

## Discussion

Two-level HLM has been empirically validated and recommended to summarize single-case data across participants (Ferron et al., 2009; Ferron et al., 2014). The use and advantages of using HLM to quantify intervention effects, with the potential of including moderators to explain heterogeneity have been documented in several

**Table 7** Power to Estimate Intervention and Moderator Effects for Two Selected SCED Studies

| | *Parameter* | King et al. (2017) | Shernoff and Kratochwill (2007) |
|---|---|---|---|
| | | Power | |
| Average baseline level | θ0 | **1.00** | **1.00** |
| Average intervention effect | θ1 | .34 | **.98** |
| Average gender moderator effect | θ2 | .19 | .14 |
| Average race moderator effect | θ3 | .22 | .39 |

*Note 1*. Power is calculated based on 1,000 replications

*Note 2*. 91 out of 1000 replications fail to converge for King et al. (2017)

*Note 3*. Power larger than .80 is indicated in bold

research papers (e.g., Baek et al., 2014; Moeyaert et al., 2014a; Moeyaert et al., 2014b; Shadish et al., 2013; Van den Noortgate & Onghena, 2007) and leading textbooks in the field (Kratochwill & Levin, 2014; Ledford & Gast, 2018). However, methodological work in this area is missing. In addition, several applied SCED studies using two-level HLM acknowledge the importance of including moderator effects to explain between participants variability. However, these applied studies lack the expertise to build a two-level model with the inclusion of moderators, and state that methodological research is needed to investigate the appropriateness of the two level HLM to include moderators (Asaro-Saddler et al., 2017; Brosnan et al., 2018; Klingbeil et al., 2017). The current study fills this gap by investigating the power of the two-level HLM approach, including moderators, to estimate intervention and moderator effects, and providing an empirical demonstration using real data. This study contributes to the literature by investigating under which conditions it might be possibly to use HLM to quantify and test for moderator effects. In addition, a newly developed point and click Shiny tool, *PowerSCED*, was presented. This user-friendly tool was developed to assists applied researchers in designing a powerful SCED study.

The results of the simulation study indicate that in general the number of participants and measurement occasions, and the size of the effect all had an impact on the power to estimate the corresponding effect. Studies with larger values for the true effects, larger number of measurement occasions and participants had higher power estimates. Consistent with past research, unit changes made at the second level of the hierarchical linear model (i.e., number of participants) had a larger effect on subsequent power estimates compared to units at the lower level (number of measurement occasions). The results comparing four models indicate that more moderators required more participants, more measurement occasions, and larger magnitudes of effects to ensure sufficient power. Studies with one moderator (nominal with two categories) needed at least 12 participants to have sufficient power to capture the intervention effect (regardless of the size of the effect), whereas the same studies not only needed at least 12 participants but also required a large moderator effect (e.g., for this simulation study, moderator effect = 3) to detect the moderator effect with sufficient power. Two binary nominal moderators required at least 20 participants to ensure sufficient power to detect the intervention effect. Whereas to detect these two binary nominal moderator effects with sufficient power, two conditions should be satisfied: at least 20 participants and large moderator effects (e.g., for this simulation study, moderator effect 1 = 3, moderator effect 2 = 1.5). When there were three moderators (two nominal and one continuous), researchers needed at least 20 participants and 40 measurements occasions to reach sufficient power to identify the intervention effect and large moderator effects. The systematic review study of Shadish and Sullivan (2011) indicated that the number of participants within a single-case study ranged from 1 to 13, and the number of observations within a participants ranged from 2 to 160. A recent meta-analysis of SCED studies (Moeyaert et al., 2019) stated that the number of participants within a SCED studies ranged from 1 to 48 with an average of 8, and that the number of measurements within a participants ranged from 6 to 68 with an average of 25. As such, the number of units (measurements and participants) needed to have sufficient power to detect true intervention and moderator effects for the models investigated in this study are not unexceptional for the field. As with any simulation study, the obtained results are limited to the conditions included in the simulation study. The current studies solely

focused on MBDs because of its popularity and its high internal and external validity. Future studies can focus on other design types such as alternating treatment designs and reversal designs. The number of participants per level of the categorical moderators was balanced in current study. An avenue for future research is to evaluate the influence of imbalanced moderators on statistical properties. Although the conditions are representative for the field, applied researchers might include a different number of measurement occasions, participants, number of moderators, and scale of moderators. In addition, different values for the true effects, and the variability might be anticipated. To address this, the simulation code was translated into a point-and-click Shiny tool, *PowerSCED*, which is freely accessible. The user can define all desired conditions and hypothesized parameter values, to evaluate whether their SCED study has sufficient power to identify true intervention and moderator effects. Researchers are encouraged to use the *PowerSCED* tool to carefully design their SCED study. This will ensure that the experiment is designed in a way that true intervention and moderator effects can be estimated with sufficient power.

**Funding** This research was supported by the Institute of Education Sciences, U.S. Department of Education, through grant R305D190022.

## Declarations

**Conflict of Interest** The authors do not have known conflict of interest to disclose.

**Disclaimer** The content is solely the responsibility of the author and does not necessarily represent the official views of the Institute of Education Sciences, or the U.S. Department of Education.

## References

Asaro-Saddler, K., Saddler, B., Moeyaert, M., & Ellis-Robinson, T. (2017). Effects of a summarizing strategy on written summaries of children with emotional and behavioral disorders. *Remedial & Special Education, 38*(2), 87–97.

Baek, E., Moeyaert, M., Petit-Bois, M., Van den Noortgate, W., Beretvas, S., & Ferron, J. (2014). The use of multilevel analysis for integrating single-case experimental design results within a study and across studies. *Neuropsychological Rehabilitation, 24*(3–4), 590–606. https://doi.org/10.1080/09602011.2013.835740.

Barlow, D. H., Nock, M. K., & Hersen, M. (2008). *Single case experimental designs: Strategies for studying behavior change* (3rd ed.). Allyn & Bacon.

Barton, E. E., Lloyd, B. P., Spriggs, A. D., & Gast, L. D. (2018). Visual analysis of graphed data. In J. Ledford & D. L. Gast (Eds.), *Single case research methodology*. Routledge.

Brosnan, J., Moeyaert, M., Brooks, K., Healy, O., Heyvaert, M., Onghena, P., & Van den Noortgate, W. (2018). Multilevel analysis of multiple-baseline data evaluating precision teaching as an intervention for improving fluency in foundational reading skills for at risk readers. *Exceptionality, 26*(3), 137–161.

Cohen, J. (1969). *Statistical power analysis for the behavioural sciences*. Academic Press.

Declercq, L., Cools, W., Beretvas, S. N., Moeyaert, M., Ferron, J. M., & Van den Noortgate, W. (2020). MultiSCED: A tool for (meta-)analyzing single-case experimental data. *Behavior Research Methods, 52*, 177–192. https://doi.org/10.3758/s13428-019-01216-2.

Ferron, J. (2005). Reversal design. In B. Everitt & D. Howell (Eds.), *Encyclopedia of behavioral statistics* (Vol. 4; pp. 1759–1760). Wiley & Sons.

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., 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., Joo, S. H., & Levin, J. R. (2017). A Monte Carlo evaluation of masked visual analysis in response-guided versus fixed-criteria multiple-baseline designs. *Journal of Applied Behavior Analysis, 50*(4), 701–716.

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*(4), 493–510.

Ferron, J., & Scott, H. (2005). Multiple baseline designs. In B. Everitt & D. Howell, (Eds.), *Encyclopedia of behavioral statistics* (Vol. 3; pp. 1306–1309). Wiley & Sons.

Gast, D. L., & Ledford, J. (2018). Combination and other designs. In J. Ledford & D. L. Gast (Eds.), *Single case research methodology*. Routledge, pp 335–364.

Gast, D. L., Lloyd, B. P., & Ledford, J. (2018). Multiple baseline and multiple probe designs. In L. Ledford & D. L. Gast (Eds.), *Single case research methodology*. Routledge, pp 239–281.

Hembry, I., Bunuan, R., Beretvas, S. N., Ferron, J. M., & Van den Noortgate, W. (2015). Estimation of a nonlinear intervention phase trajectory for multiple-baseline design data. *Journal of Experimental Education, 83*(4), 514–546.

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. https://doi.org/10.1016/j.ridd.2011.10.010

Horner, R. H., & Odom, S. L. (2014). Constructing single-case research designs: Logic and options. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances*. American Psychological Association, pp 27–51.

Jamshidi, L., Heyvaert, M., Declercq, L., Fernández-Castilla, B., Ferron, J. M., Moeyaert, M., Beretvas, N. S., Onghena, P., & Van den Noortgate, W. (2020). A systematic review of single-case experimental design meta-analyses: Characteristics of study designs, data and analyses. *Evidence-Based Communication Assessment & Intervention.*

Kazdin, A. E. (2011). *Single-case research designs: Methods for clinical and applied settings* (2nd ed.). Oxford University Press.

King, B., Radley, K. C., Jenson, W. R., & O'Neill, R. E. (2017). On-task in a box: An evaluation of a package-ready intervention for increasing levels of on-task behavior and academic performance. *School Psychology Quarterly, 32*(3), 306–319.

Klingbeil, D., Moeyaert, M., Archerm, C., Chimnoza, T. M., & Zwolski, S. A. (2017). Efficacy of peer-mediated incremental rehearsal for English Language Learners. *School Psychology Review, 46*(1), 122–140.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). *Single-case design technical documentation*. What Works Clearinghouse. http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf

Kratochwill, T. R., & Levin, J. R. (Eds.). (2014). *Single-case intervention research: Methodological and statistical advances*. *American Psychological Association.* https://doi.org/10.1037/14376-000.

Kratochwill, T. R., Levin, J. R., Horner, R. H., & Swoboda, C. M. (2014). Visual analysis of single-case intervention research: Conceptual and methodological issues. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances*. American Psychological Association, pp. 91–125.

Ledford, J. R., & Gast, D. L. (Eds.). (2018). Single case research methodology (3rd ed.). Routledge.

McKnight, S. D., McKean, J. W., & Huitema, B. E. (2000). A double bootstrap method to analyze linear models with autoregressive error terms. *Psychological Methods, 5*, 87–101.

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

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

Moeyaert, M., Ugille, M., Ferron, J., Beretvas, S., & Van den Noortgate, W. (2014b). The influence of the design matrix on treatment effect estimates in the quantitative analyses of single-case experimental design research. *Behavior Modification, 38*(5), 665–704. https://doi.org/10.1177/0145445514535243.

Moeyaert, M., Akhmedjanova, D., Ferron, J., Beretvas, S. N., & Van den Noortgate, W. (2020). Effect size estimation for combined single-case experimental designs. *Evidence-Based Communication Assessment & Intervention, 14*(1–2), 28–51.

Moeyaert M, Yang P, Xu X, & Kim E. (2021a). Characteristics of moderators in meta-analyses of single-case experimental design studies. Behavior Modification. https://doi.org/10.1177/01454455211002111

Moeyaert, M., Klingbeil, D., Rodabaugh, E. & Turan, M. (2021b). Three-level meta-analysis of single-case data regarding the effects of peer tutoring on academic and social-behavioral outcomes for at-risk students and students with disabilities. Remedial and Special Education, 42(2), 94-106. https://doi.org/10.1177/0741932519855079

Parker, R. I., Vannest, K. J., & Brown, L. (2009). The improvement rate difference for single-case research. *Exceptional Children, 75*(2), 135–150.

Parker, R. I., Vannest, K. J., Davis, J. L., & Sauber, S. B. (2011). Combining nonoverlap and trend for single-case research: Tau. *Behavior Therapy, 42*(2), 284–299.

Petit-Bois, M., Baek, E. K., Van den Noortgate, W., Beretvas, S. N., & Ferron, J. M. (2016). The consequences of modeling autocorrelation when synthesizing single-case studies using a three-level model. *Behavior Research Methods, 48*(2), 803–812.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Houghton-Mifflin.

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.

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.

Shernoff, E. S., & Kratochwill, T. R. (2007). Transporting an evidence-based classroom management program for preschoolers with disruptive behavior problems to a school: An analysis of implementation, outcomes, and contextual variables. *School Psychology Quarterly, 22*(3), 449–472.

Tarlow, K. R. (2017). An improved rank correlation effect size statistic for single-case designs: Baseline corrected Tau. *Behavior Modification, 41*(4), 427–467.

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

Van den Noortgate, W., & Onghena, P. (2003b). Multilevel meta-analysis: A comparison with traditional meta-analytical procedures. *Educational & Psychological Measurement, 63*(5), 765–790.

Van den Noortgate, W., & Onghena, P. (2007). The aggregation of single-case results using hierarchical linear models. *The Behavior Analyst Today, 8*, 196–209.

Van den Noortgate, W., & Onghena, P. (2008). A multilevel meta-analysis of single-subject experimental design studies. *Evidence-Based Communication Assessment & Intervention, 2*(3), 142–151. https://doi.org/10.1080/17489530802505362.

What Works Clearinghouse. (2020). *What Works Clearinghouse standards handbook* (Version 4.1). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. https://ies.ed.gov/ncee/wwc/handbooks

Wolery, M., Gast, D. L., & Ledford, J. (2018). Comparative and other designs. In J. Ledford & D. L. Gast (Eds.), *Single case research methodology*. Routledge, pp 283–328.

. Xu, X., Moeyaert, M., & Yang, Y. (2021). PowerSCED (Version 1.0) [Web application]. https://xinyunxu.shinyapps.io/PowerSCED/_w_8b4d5ac0