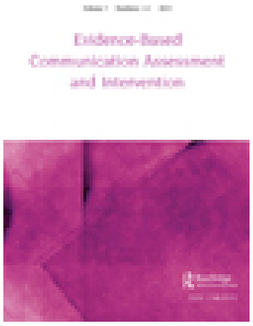
Taylor & Francis
Taylor & Francis Group



Evidence-Based Communication Assessment and Intervention

ISSN: 1748-9539 (Print) 1748-9547 (Online) Journal homepage: https://www.tandfonline.com/loi/tebc20

Effect size estimation for combined single-case experimental designs

Mariola Moeyaert, Diana Akhmedjanova, John Ferron, S. Natasha Beretvas & Wim Van den Noortgate

To cite this article: Mariola Moeyaert, Diana Akhmedjanova, John Ferron, S. Natasha Beretvas & Wim Van den Noortgate (2020): Effect size estimation for combined single-case experimental designs, Evidence-Based Communication Assessment and Intervention

To link to this article: https://doi.org/10.1080/17489539.2020.1747146

Published online: 30 Apr 2020.

Submit your article to this journal

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=tebc20

Evidence-Based Communication Assessment and Intervention, 2020
https://doi.org/10.1080/17489539.2020.1747146

Routledge
Taylor & Francis Group

Check for updates

EBP Advancement Corner

Effect size estimation for combined single-case experimental designs

Mariola Moeyaert[1], Diana Akhmedjanova[1], John Ferron[2], S. Natasha Beretvas[3] & Wim Van den Noortgate[4]

[1]Department of Educational and Counseling Psychology, University at Albany-SUNY, Albany, NY, USA; [2]Department of Educational Measurement and Research, University of South Florida, Tampa, FL, USA; [3]Department of Educational Psychology, University of Texas, Austin, TX, USA; [4]Faculty of Psychology and Educational Sciences & Imec-itec, KU Leuven, Leuven, Belgium

Abstract

The methodology of single-case experimental designs (SCED) has been expanding its efforts toward rigorous design tactics to address a variety of research questions related to intervention effectiveness. Effect size indicators appropriate to quantify the magnitude and the direction of interventions have been recommended and intensively studied for the major SCED design tactics, such as reversal designs, multiple-baseline designs across participants, and alternating treatment designs. In order to address complex and more sophisticated research questions, two or more different single-case design tactics can be merged (i.e., "combined SCEDs"). The two most common combined SCEDs are (a) a combination of a multiple-baseline design across participants with an embedded ABAB reversal design, and (b) a combination of a multiple-baseline design across participants with an embedded alternating treatment design. While these combined designs have the potential to address complex research questions and demonstrate functional relations, the development and use of proper effect size indicators lag behind and remain unexplored. Therefore, this study probes into the quantitative analysis of combined SCEDs using regression-based effect size estimates and two-level hierarchical linear modeling. This study is the first demonstration of effect size estimation for combined designs.

Keywords: *Combined designs; effect size; hierarchical linear modeling; regression models; single-case experimental design.*

Single-case experimental designs (SCEDs) are rigorous experimental designs that have been applied in a variety of fields (e.g., biomedical research, language and speech therapy, behavior modification, school psychology, counseling psychology, physical therapy, special education, and neuropsychological rehabilitation) to evaluate the efficacy and effectiveness of interventions (Kennedy, 2005; Kratochwill et al., 2014; Moeyaert, Ferron, et al., 2014). In SCEDs, a case (one unit [e.g., participant], or an aggregate unit such as a class) is measured repeatedly across time during conditions (e.g., baseline and intervention condition or multiple intervention conditions). Data from different conditions are compared to evaluate the efficacy or effectiveness of one or multiple interventions. The basic question examined using SCEDs is whether there is evidence for a functional relation between the systematic manipulation of an independent variable (i.e., the conditions) and its consistent effect on a dependent variable (i.e., the target behavior) (Kratochwill et al., 2010; Kratochwill & Levin, 2014; J. Ledford et al., 2018).

Valid and reliable structured visual analysis techniques (J. Ferron & Jones, 2006;

For correspondence: Mariola Moeyaert, School of Education, Department of Educational and Counseling Psychology, Division of Educational Psychology & Methodology, The University at Albany - SUNY, 1400 Washington Ave, Albany, NY 12222. E-mail: mmoeyaert@albany.edu

© 2020 Informa UK Limited, trading as Taylor & Francis Group

Kratochwill et al., 2010) have been developed for interpreting SCED results and are widespread. Visual analysis has a rich history and is strongly embedded in the field of SCEDs. It is considered to be a valid approach for identifying "weak", "moderate", or "strong" evidence for a causal relationship between an independent and dependent variables by evaluating data using six steps described by Kratochwill et al. (2010). Following the technical documentation of the *What Works Clearinghouse (WWC) Standards for Design and Analysis of SCEDs* (Kratochwill et al., 2010), the field is now moving toward estimating effect size indicators to supplement and support the visual analysis results. Efforts have been made to develop effect size estimates for "single" SCEDs such as the alternating treatment design, multiple-baseline design, and ABAB reversal design (e.g., Lenz, 2013; Maggin et al., 2011; Manolov & Solanas, 2013; Moeyaert, Ugille, Ferron, Beretvas, et al., 2014; Moeyaert, Ugille, Ferron, Onghena, et al., 2014; Parker, Vannest, & Davis, 2011; Parker et al., 2014; Shadish et al., 2008, 2014; Swaminathan et al., 2010; Wolery et al., 2010). However, the formulation of these effect size indicators for "combined" SCEDs is not yet fully developed. This study is timely, especially given the potential of these types of designs to answer rich research questions and to make internally and externally more valid inferences about the efficacy or effectiveness of an intervention.

Combined single-case designs

Shadish and Sullivan (2011) conducted a review of SCED studies published in 2008 to review their design and data characteristics. Their search resulted in 809 unique SCED studies, 73.1% of which consisted of "single" designs: 54.3% were Multiple-Baseline Designs (MBD) across participants; 8.2% represented Withdrawal and Reversal Designs (WRD, such as ABAB reversal designs); 8.0% were Alternating Treatment Designs (ATDs); and 2.6% were Changing Criterion Designs (CC). The authors found that a proportion of SCEDs (26.9%) do not use a "single" design, but rather a design that combines characteristics of two or more "single" SCED designs – so-called "combined SCEDs" (J. Ledford & Gast, 2018). Specifically, the combination of MBD + WRD appeared to be the most popular one (12.0%), followed by the combination of MBD + ATD (9.9%).

Combined or combination SCEDs (J. Ledford & Gast, 2018) offer three major advantages compared to single SCEDs. First, they allow assessment of multiple research questions. For example, Trottier et al. (2011) looked at the functional relation between peer-tutoring interventions and the number of spontaneous appropriate communicative acts generated by students with autism spectrum disorder (ASD) as the main focus of their study. The use of a combined SCED let the researchers examine whether normally developing peers could independently teach children with ASD to use speech-generating devices or whether the typically developing peers had to first be taught how to instruct the children with ASD. As a result, this combined design study allowed the researchers to evaluate two different interventions simultaneously: (a) teaching typically developing peers to give timely prompts to children with ASD to use the device; and (b) letting typically developing peers teach children with ASD to use the device (Trottier et al., 2011). Additionally, the two interventions were alternated for each child, and the interventions were staggered across participants (n = 2), resulting in an MBD + ATD combined design.

Second, a combined SCED allows for more evaluations of the effectiveness of

the treatment as more replications are present. For example, the MBD + WRD combined design allows for replication of a treatment effect after removing and reintroducing the treatment within a participant as well as across participants, taking into account different start times for the treatment. In case of the MBD + ATD combined design, the replication of alternating treatments can be seen both within each participant and across participants at different points in time. The replication effects can be identified both within and across participants. Replication is a central theme in SCED studies (Kratochwill et al., 2010) because it enhances the external validity of the resulting conclusions. Indeed, there is additional documentation of the effect at more points in time and more replications within one case.

Third, due to the dynamic nature of combined designs, they grant an opportunity to modify pure SCEDs by adding design elements in the middle of the study. For instance, Kelley et al. (2002) initially used an MBD to investigate the effectiveness of competing reinforcement schedules on functional communication (Figure 1). However, the data demonstrated problems. The disruptive behaviors for two out of the three participants were not decreasing; as a result, the authors slightly changed the condition from Functional Communication Training (FCT) without extinction to FCT with extinction, ensuring treatment fidelity for all the other steps in the study. In this way, the introduction of the ABAB allowed the study to continue and provided an opportunity to address the core research question.

The analysis of the majority of the combined design studies typically relies on visual analyses and non-overlap indices to identify and make inferences about the intervention effects (Chung & Cannella-Malone, 2010; Jason & Frasure, 1979; Matson & Keyes, 1990; Trottier et al., 2011). For example, Lindberg et al. (1999) used an MBD + WRD combined design study to evaluate the effects of manipulation and reinforcement on self-injurious behaviors of two participants, solely relying on visual analysis. Another combined SCED study, MBD + ATD (Trottier et al., 2011), reported the results of the effectiveness of peer-tutoring on the use of speech-generating devices for students with autism in social game routines using visual analysis and the Percentage of Non-Overlapping Data index (PND; Schlosser et al., 2008; Scruggs et al., 1987)). Relying on visual analysis and non-overlap indices is unfortunate because the opportunity is lost to precisely address additional questions through quantitative summaries (e.g., What is the magnitude of the intervention effect? To what extent is the intervention immediately effective? To what extent does the intervention remain effective over time? Are all the participants benefiting equally from the intervention?). While visual analysis and non-overlap indices provide an initial indication of effectiveness of an intervention, effect size indices are needed to provide additional information through quantitative synthesis. Effect size indicators can be used to quantify the magnitude of intervention effectiveness at multiple points in time both for each participant and across participants. In addition, effect size estimates are supplemented with a standard error that reflects precision for the individual estimate and which can be used as a weight for quantitative summaries or analyses (i.e., multilevel meta-analysis; Moeyaert, 2019). Therefore, in this article, we are breaking new ground by applying the effect size logic to quantify intervention effectiveness for combined SCEDs. The effect size estimates will provide a more comprehensive picture regarding intervention effects by taking into account the design complexity of combined SCEDs, and they can be used in meta-

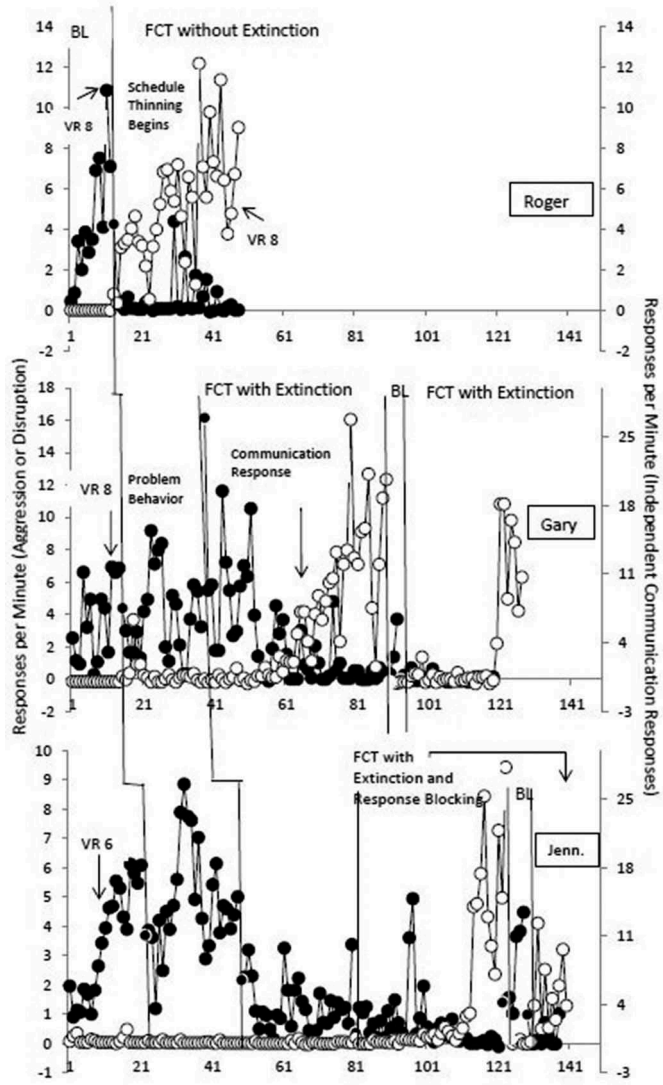


Figure 1. An example of modifying the multiple baseline design by adding a phase change reversal. Frequency of target behaviors for three participants. Adapted from "The Effects of Competing Reinforcement Schedules on the Acquisition of Functional Communication," by M. E. Kelley, D. C. Lerman, and C. M. Van Camp, 2002, Journal of Applied Behavior Analysis, 35(1), p. 62.

analyses to assess generalizability across interventions and outcome variables.

Previous research has focused on the coding schemes and synthesis of results for each of the "single" SCEDs, including the simple AB phase design, the MBD across participants, WRD (ABAB), and ATDs (Moeyaert, Ugille, Ferron, Onghena, et al.,

2014; Shadish, Kyse et al., 2013). Researchers have not investigated (1) coding and effect size estimation for combined SCEDs, and (2) meta-analysis of studies involving combined SCEDs. Due to the lack of methodology to quantify combined SCEDs, these studies tend to be simplified or excluded from meta-analyses, which contributes to biased effect size estimates and/or publication bias (e.g., Kokina & Kern, 2010; Wang et al., 2013). Therefore, we focus on how to quantify treatment effects for combined designs. Thus, the purpose of this study is to illustrate effect size estimation for combined designs using real data. In particular, we will focus on the MBD + WRD combined designs (=45.97%) and the MBD + ATD combined designs (=37.91%) as they are the two most popular classes of combined SCEDs: 83.38% of the combined SCEDs (Shadish & Sullivan, 2011).

METHOD

We identified combined design studies and then randomly selected one MBD + WRD and one MBD + ATD study. Combined SCEDs were identified by examining primary studies from four meta-analyses of SCEDs (Heyvaert et al., 2014; Kokina & Kern, 2010; Moeyaert et al., 2019; Shogren et al., 2004) and 20 primary studies that evaluated reading fluency interventions. These meta-analyses and primary SCED studies were chosen because the first author had access to raw data. The meta-analysis of Heyvaert et al. (2014) included 59 studies of which 11 studies (i.e., 18.64%) were combined SCEDs. The review by Kokina and Kern (2010) consisted of 18 SCEDs of which only four (i.e., 22.22%) were combined SCEDs. The peer-tutoring meta-analysis by Moeyaert et al. (2019) included 65 studies and contained nine combined SCEDs (i.e., 13.85%). The last meta-analysis (Shogren et al., 2004) had 13 SCED studies and two of them (15.38%) were combined SCEDs. Finally, of the 20 primary studies that examined reading fluency interventions, seven (i.e., 35%) were combined SCEDs. Thus, a substantial proportion of reviewed studies was combined SCEDs, a finding that is consistent with the review of Shadish and Sullivan (2011). The full list of the 33 combined design studies from the meta-analyses that we reviewed is available from the first author upon request. Of these combined designs, the combinations MBD + WRD (i.e., 58.82%, 20 studies) and MBD + ATD (i.e., 23.52%, eight studies) were the most popular. This also supports the results from the study of Shadish and Sullivan (2011) and our decision to focus on these two classes of combined SCEDs in this study.

One study per combined SCED type was randomly selected from the set to demonstrate the coding of the design matrix and estimation of the effect sizes. The design matrix gives an overview of the overall data structure and includes all variables (e.g., participant identifier, the dependent variable, the independent variables) together with scores assigned to these variables. All variables needed to estimate the effect sizes of interest should be reflected in the design matrix. For more information about the design matrix for SCEDs, see Moeyaert, Ugille, Ferron, Beretvas et al. (2014). However, other studies from the selection could also have been chosen. Raw data for the dependent variable in SCEDs are traditionally graphically displayed as can be seen in Figure 2 (MBD + WRD) and Figure 3 (MBD + ATD). As a result, researchers can retrieve raw data from the graphical displays in primary studies. We used WebPlotDigitizer (Rohatgi, 2011) to retrieve raw data. The raw data represent the measures of the dependent variable over time. The dependent variable (i.e., targeted behavior) together with other variables (i.e., phase and time indicators) that

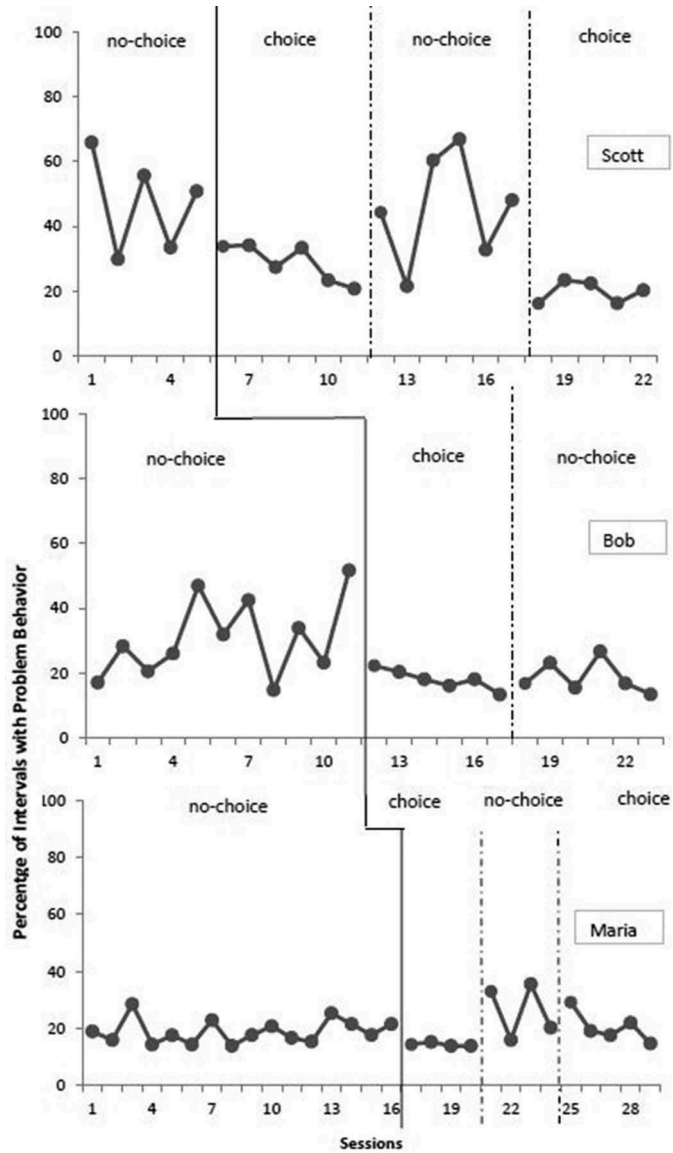


Figure 2. An example of the mixed design: MBD + PCR. Percentage of intervals with problem behaviors for three participants. Adapted from "The Effects of Choice-making on the Problem Behaviors of High School Students with Intellectual Disabilities," by S. Seybert, G. Dunlap, and J. Ferro, 1996, Journal of Behavior Education, 6 (1), p. 58.

are needed to conduct the statistical analysis are part of the design matrix. The design matrix needed for effect size estimation of the combined designs is displayed in Tables 1 and 4 and will be discussed later. For more information about the data retrieval process, see Moeyaert, Maggin, et al. (2016). The raw data from Figures 2 and 3 can be found in the

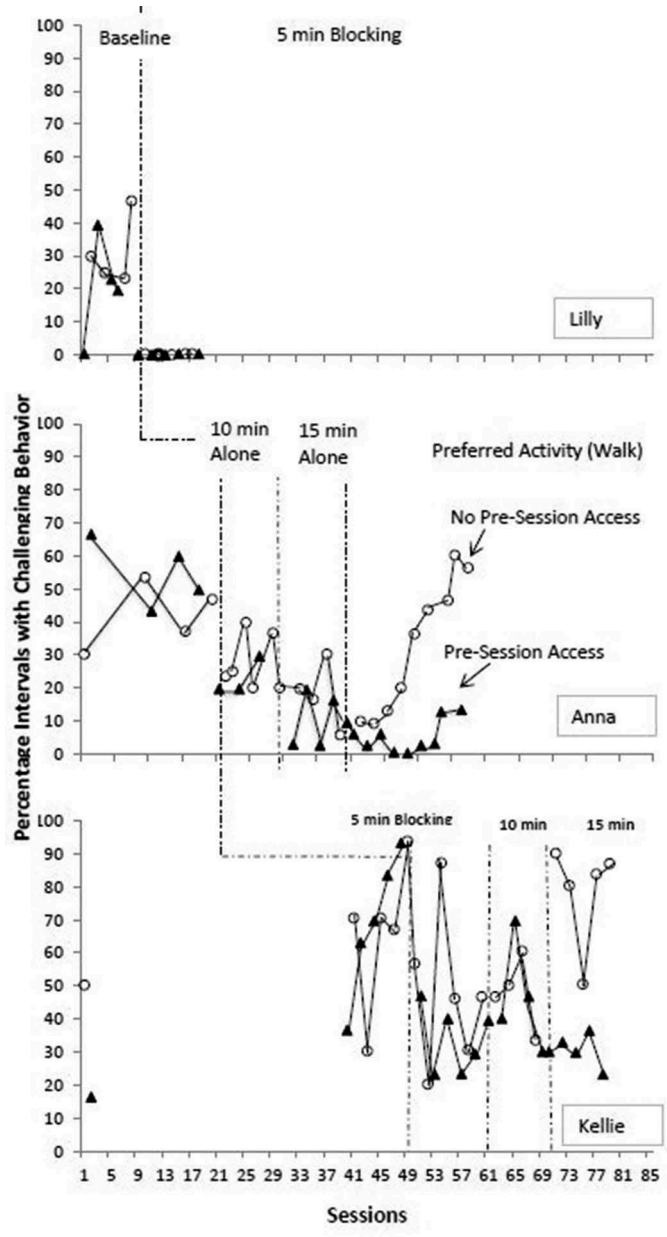


Figure 3. An example of the Mixed Design: MBD + ATD. Percentage intervals with challenging behavior for three participants. Adapted from "The Effects of Presession Manipulations on Automatically Maintained Challenging Behavior and Task Responding," by Y.-C. Chung, and H. I. Cannella-Malone, 2010, Behavior Modification, 34(6), p. 493.

supplement to this article (together with the SAS codes that can be used for the analyses) to facilitate replication of the analyses demonstrated in this study, using the same data sets.

Table 1. Design matrix for Case 1 (i.e., Scott) – Seybert et al. (1996)

Case	Session	Outcome	A1B1	B1A2	A2B2
1	1	65.92	0	0	0
1	2	29.89	0	0	0
1	3	55.71	0	0	0
1	4	33.46	0	0	0
1	5	50.84	0	0	0
1	6	33.82	1	0	0
1	7	34.15	1	0	0
1	8	27.39	1	0	0
1	9	33.36	1	0	0
1	10	23.35	1	0	0
1	11	20.75	1	0	0
1	12	44.32	1	1	0
1	13	21.51	1	1	0
1	14	60.35	1	1	0
1	15	66.91	1	1	0
1	16	32.76	1	1	0
1	17	48.10	1	1	0
1	18	16.19	1	1	1
1	19	23.37	1	1	1
1	20	22.43	1	1	1
1	21	16.24	1	1	1
1	22	20.29	1	1	1

RESULTS

Effect sizes are used as a complement to visual analysis in primary studies and can be used for between-study comparison of treatment effects and for meta-analytic purposes. Visual analysis has been well documented by Kratochwill et al. (2010), whereas the focus of the current study is on the quantitative summary of combined SCEDs. The analyses in the empirical illustration sections are performed using SAS software, Version 9.4 (© SAS Institute Inc.) SAS codes are available in the supplement to this article.

Multiple-baseline design – Withdrawal or reversal design

To demonstrate the effect size estimation for the first class of combined SCEDs, we selected the study of Seybert et al. (1996). Seybert et al. (1996) investigated the differences in problem and on-task behaviors in choice and no-choice conditions of three independent participants with intellectual disabilities. In the choice condition, participants were given a choice of the domestic task to do. In contrast, in the no-choice condition, participants were assigned to do

Table 2. Results of ordinary least squares analysis and Empirical Bayes analysis per participant

Case	Parameter	OLS Estimate (*SE*)	Estimate (*Standard error of prediction*)
Scott	$\hat{\beta}_{01}$	61.31 (6.90)	57.74 (11.87)
	$\hat{\beta}_{11}$	−24.28 (9.34)	−19.30 (-)
	$\hat{\beta}_{21}$	20.73 (8.91)	18.02 (10.35)
	$\hat{\beta}_{31}$	−40.01 (9.34)	−37.38 (15.50)
Bob	$\hat{\beta}_{02}$	38.20 (4.99)	36.37 (11.77)
	$\hat{\beta}_{12}$	−22.47 (8.39)	−19.30 (-)
	$\hat{\beta}_{22}$	1.31 (9.55)	2.11 (10.32)
Maria	$\hat{\beta}_{03}$	16.53 (2.97)	18.90 (11.77)
	$\hat{\beta}_{13}$	−12.82 (6.64)	−19.30 (-)
	$\hat{\beta}_{23}$	26.99 (8.40)	29.85 (10.44)
	$\hat{\beta}_{33}$	−10.90 (7.97)	−10.98 (15.50)

Table 3. Results of two-level analysis across participants

Parameter		**Estimate (*SE*)**	***t***	***p***
Fixed Effects				
Baseline level A1	$\hat{\theta}_0$	37.67 (11.74)	3.21	.082
Change in level A1 – B1	$\hat{\theta}_1$	−19.30 (4.59)	−4.21	<.001
Change in level B1 – A2	$\hat{\theta}_2$	16.66 (10.26)	1.62	.227
Change in level A2 – B2	$\hat{\theta}_3$	−24.18 (15.76)	−1.53	.367
Random Effects		**Estimate (*SE*)**	***z***	***p***
Baseline level A1	$\hat{\sigma}^2_{u_0}$	391.51 (406.93)	0.96	.168
Change in level A1 – B1	$\hat{\sigma}^2_{u_1}$	0 (/)	/	/
Change in level B1 – A2	$\hat{\sigma}^2_{u_2}$	236.77 (291.24)	0.81	.208
Change in level A2 – B2	$\hat{\sigma}^2_{u_3}$	414.59 (701.37)	0.59	.277
Within-case variance	$\hat{\sigma}^2_e$	207.75 (36.40)	5.71	<.0001

Table 4. Design matrix for Case 1 – Data retrieved from Chung and Cannella-Malone (2010)

Case	**Session**	**Outcome**	Treatment_1	Treatment_2
1	1	0.27933	0	0
1	2	29.88827	0	0
1	3	39.38547	0	0
1	4	24.86034	0	0
1	5	22.90503	0	0
1	6	19.55307	0	0
1	7	23.18436	0	0
1	8	46.64804	0	0
1	9	0	1	0
1	10	0.27933	0	1
1	11	0	1	0
1	12	0	0	1
1	13	0	1	0
1	14	0	0	1
1	15	0.27933	1	0
1	16	0.27933	0	1
1	17	0.27933	0	1
1	18	0.27933	1	0

a certain domestic task. The outcome variable reflected the percentage of problem behaviors and task engagement in the choice versus no-choice conditions. The data were recorded using the 15-s partial interval recording: that is, only the five last seconds was recorded per each 15-s interval. Data points per participant ranged from n = 22 (Scott) to n = 29 (Maria). Seybert et al. (1996) used the combination of the MBD + WRD to investigate the effectiveness of choice-making on problem behavior. A graphical display is given in Figure 2. Seybert et al. (1996) claimed that the MBD + WRD allowed them to provide further evidence for the changes in the treatment

phase as a result of manipulating the independent variable – choice versus no-choice conditions. The inter-rater observer percent agreement ranged from 81% to 99% for occurrence and nonoccurrence of problem behaviors. Seybert et al. (1996) analyzed the data using visual analysis techniques, and the results were reported as percentages of intervals with problem behaviors. This combined SCED has the potential to demonstrate a functional relation between the choice-making condition and problem behavior as the effectiveness of the treatment can be evaluated at three or more different points in time. In addition, most of the phases included at least five measurements (one choice and one no-choice condition for Maria included only four measurements). The MBD embedded in the combined design meets the WWC design standards as it includes at least three potential demonstrations of treatment effectiveness across at least three different points in time. The WRD embedded in the combined design meets basic replications standards for Scott and Maria whereas this is not the case for Bob. There appears to be a non-effect for the withdrawal of the treatment. In addition, the WRD for Bob does not meet the WWC design standards as there are only two potential demonstrations of treatment effectiveness. According to Gast et al. (2018) this prohibits the conclusion that a functional relation is present for Bob. Notwithstanding of this non-effect and lack of experimental control for Bob, effect size estimation for this combined design can still be meaningful. Researchers might be interested in quantifying the size of the effect, and this quantification can be used to confirm the results based on the visual analysis. This effect size estimate can be used afterward for meta-analytic purposes. We focused on estimating regression-based effect size estimates for the occurrence of problem behaviors in choice-making conditions for three participants with intellectual disabilities. The statistical model and empirical illustration are discussed in the following sections.

Statistical model Step 1: single-level analysis. The single-level analysis can also be called an *individual analysis* as it involves a case-by-case evaluation of treatment effectiveness. Here, we are interested in demonstrating the effectiveness of a treatment at different points in time within participants. In the simplest scenario, the results are an estimate of change in levels between baseline and treatment phases for each participant separately. In other words: *"Is there evidence for change in level between adjacent phases?"* In this particular scenario, the design matrix contains dummy-coded variables indicating the specific phase to which a measurement belongs (see Table 1). We chose the following notation to distinguish between the consecutive phases: A1 and A2 indicate, respectively, the first and the second baseline phase, and B1 and B2 denote the first and the second treatment phase. For the ABAB phase design, three dummy variables, *A1B1*, *B1A2*, and *A2B2* are coded as suggested by Moeyaert, Ugille, Ferron, Beretvas, et al. (2014) and Shadish, Kyse, et al. (2013). *A1B1* = 1 for all the measurement occasions after the first baseline phase, *B1A2* = 1 for all the measurement occasions after the first treatment phase and *A2B2* equals 1 during the last treatment phase (see Table 1). In order to predict the outcome score at the i*th* measurement occasion, the following multiple regression equation can be used and parameters can be estimated using Ordinary Least Squares (i.e., OLS):

$$Y_i = \beta_0 + \beta_1 A1B1_i + \beta_2 B1A2_i + \beta_3 A2B2_i + e_i \text{ with } e_i \sim N\left(0, \sigma_e^2\right) \quad (1)$$

When all three dummy-coded variables equal zero (i.e., $A1B1 = B1A2 = A2B2 = 0$), then the indicated phase is the first baseline phase (β_0). Each dummy variable represents the change from an earlier to its adjacent phase. Thus, for example, *B1A2* refers to the change in level from B1 to A2 (i.e., difference in level between Treatment 1 and Baseline 2). An extension here could be to investigate whether there are changes in linear (Moeyaert, Ugille, Ferron, Beretvas, et al., 2014) or non-linear trends (Hembry et al., 2015) or changes in variance of scores between adjacent phases (Baek & Ferron, 2013).

Statistical model Step 2: two-level analysis. The two-level analysis involves an aggregate estimate of the treatment effectiveness across participants. Here, we are investigating the replication of the treatment effect across participants (within the same study), in addition to the replication of the treatment effect within participants. As a consequence, more generalized conclusions can be made, which strengthens the external validity of the inferences. In addition, variability in effectiveness of the treatment between participants can be quantified. One way to perform this analysis is to conduct a two-level analysis, which takes the hierarchical nature of the data into account; namely, measurements are nested within each of multiple cases.

The coefficients from the first level: β_{0j}, β_{1j}, β_{2j}, and β_{3j}, can be modeled as varying at the second (participant) level. By fitting this multilevel model, overall average changes in level from one phase to another can be obtained in addition to how individual participants deviate from that overall change. The level 1 and level 2 equations are presented in Equations (2) and (3):

Level 1:

$$Y_{ij} = \beta_{0j} + \beta_{1j} A1B1_{ij} + \beta_{2j} B1A2_{ij} + \beta_{3j} A2B2_{ij} + e_{ij} \text{ with } e_{ij} \sim N\left(0, \sigma_e^2\right) \text{ with } e_{ij} \sim N\left(0, \sigma_e^2\right) \quad (2)$$

Level 2:

$$\begin{cases} \beta_{0j} = \theta_{00} + u_{0j} \\ \beta_{1j} = \theta_{10} + u_{1j} \\ \beta_{2j} = \theta_{20} + u_{2j} \\ \beta_{3j} = \theta_{30} + u_{3j} \end{cases} \text{with} \begin{bmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \\ u_{3j} \end{bmatrix}$$

$$N \sim \left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{u_0}^2 & \sigma_{u_0 u_1} & \sigma_{u_0 u_2} & \sigma_{u_0 u_3} \\ \sigma_{u_1 u_0} & \sigma_{u_1}^2 & \sigma_{u_1 u_2} & \sigma_{u_1 u_3} \\ \sigma_{u_2 u_0} & \sigma_{u_2 u_1} & \sigma_{u_2}^2 & \sigma_{u_2 u_3} \\ \sigma_{u_3 u_0} & \sigma_{u_3 u_1} & \sigma_{u_3 u_2} & \sigma_{u_3}^2 \end{bmatrix} \right) \quad (3)$$

The first line in Equation (3) indicates that the baseline level for participant j is modeled as a function of an average baseline level, θ_{00}, plus a random deviation from this mean, u_{0j}. The subsequent equations describe the average change in level between A1 and B1 (θ_{10}), change in level between B1 and A2 (θ_{20}), and change in level between A2 and B2 (θ_{30})phases, respectively. The variability in baseline level (i.e., $\sigma_{u_0}^2$) and variability in changes in levels (i.e., $\sigma_{u_1}^2$, $\sigma_{u_2}^2$ and $\sigma_{u_3}^2$) are captured by estimating the variance/covariance matrix.

Empirical illustration. We use the Seybert et al. (1996) study for the empirical illustration of the single-level (individual) and two-level (average) effect size estimates for the MBD + WRD design. Seybert et al. (1996) investigated the effects of choice-making on the problem behaviors of three high school students with intellectual disabilities. In this example, we are looking

only at the outcome variable of occurrence and nonoccurrence of problem behaviors within choice and no-choice conditions. The start of the intervention was staggered across the three participants, and two baseline conditions (i.e., no-choice – denoted as A1 and A2 in Figure 4) are interrupted by treatment conditions (i.e., choice – denoted as B1 and B2 in Figure 4). Participant 2 (i.e., Bob) has no second treatment phase as the

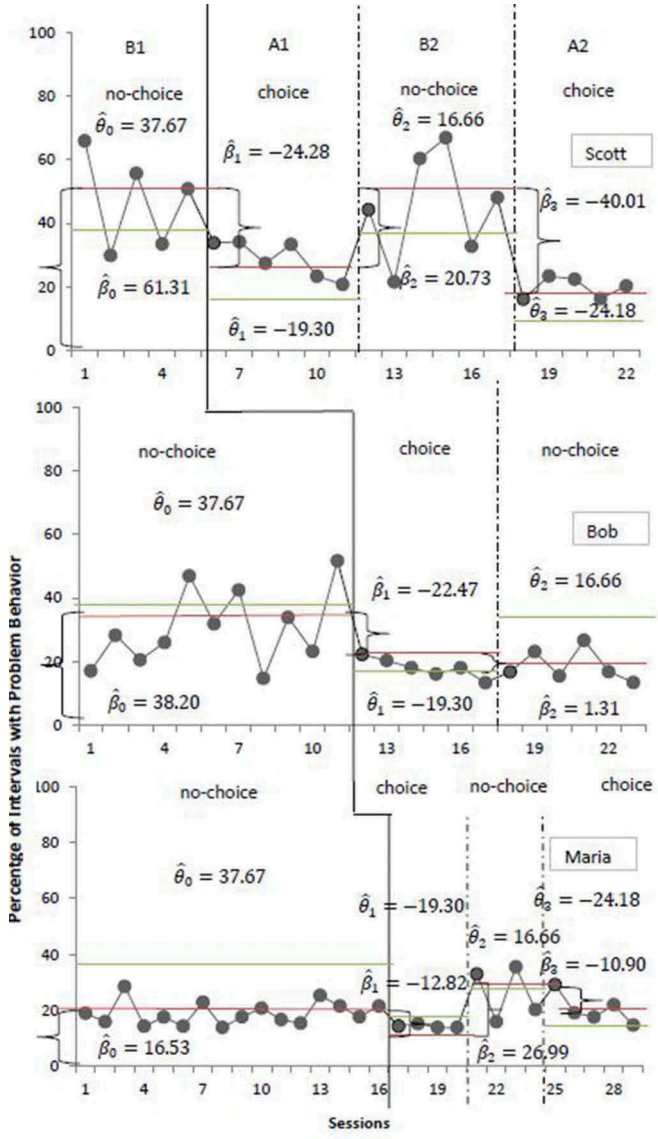


Figure 4. Estimated parameters for each participant across phases. Note: The lines indicate case-specific and study-specific estimates.

problem behavior remained low when the treatment was removed (phase A2). The graphical presentation of the data is given in Figure 2. The coding of the design matrix for participant 1 (i.e., Scott) in accordance with the mathematical model presented in Equation (1) can be found in Table 1 (the same coding is applied for the other cases). The SAS code to run the analyses is available as a supplement to this article.

The output of the single-level analysis is presented in Table 2, and the visual presentation of the estimated parameters is provided in Figure 4. From the single-level analysis, we can conclude that there is a demonstration of treatment effectiveness at three different points in time for Case 1 (i.e., Scott). When the choice-making intervention is introduced, we see a significant drop in problem behavior [$\hat{\beta}_{11} = -24.28,\ t(25) = -2.60,\ p = .018$ and $\hat{\beta}_{31} = -40.01, t(25) = -4.28, p = .032$]. When the choice-making intervention is removed, we see a significant increase in problem behavior [$\hat{\beta}_{21} = 20.73, t(25) = 2.33, p = .032;$]. For Case 2 (i.e., Bob) and Case 3 (i.e., Maria), there was only one demonstration of significant treatment effectiveness [Case 2: $\hat{\beta}_{12} = -22.47, t(20) = -2.68,\ p = .015$, and Case 3: $\hat{\beta}_{23} = 26.99, t(25) = 3.21 p = .004$]. According to the WWC design standards (Kratochwill et al., 2010), the choice-making intervention was only effective for Scott as three demonstrations of treatment effectiveness at three different points in time are required to demonstrate a causal relationship between the introduction of the treatment and the change in outcome score.

The two-level analysis was conducted to estimate the overall baseline level and changes in level between subsequent phases across the three cases in addition to between-case variability in these estimates. The two-level analysis enhances the generalizability of treatment effectiveness beyond the cases under investigation. For didactic purposes (allowing visual presentation of the estimated coefficients, Figure 4), a small dataset with only three cases is used. In order to run a two-level analysis and obtain generalizable estimates, it is suggested to use a larger dataset, including more than three cases. The results indicate that the choice-making intervention succeeded in reducing the problem behavior and large effect size estimates were obtained for the change in level between A1 and B1 and A2 and B2 [$\hat{\theta}_{10} = -19.30,\ t(66) = -4.21,\ p < .001;\ \hat{\theta}_{30} = -24.18, t(1) = -1.53, p = .367$]. However, only one estimate ($\hat{\theta}_{10}$) is statistically significant ($p < .05$).

An additional advantage of using the two-level analysis is that the between-case variance in treatment effect estimates can be estimated. Most variability was found in the estimate of the between-case variance for the change in level between A2 and B2 (Table 3, random effects). The results of the single-level and two-level analyses are visually presented in Figure 4.

Another advantage of using the two-level analysis is that empirical Bayes estimates of the case-specific parameters can be obtained. The empirical Bayes estimate can be viewed as a fully Bayesian approach that uses information of the full dataset to build prior distributions (Shadish, Rindskopf, et al., 2013). Therefore, the empirical Bayes estimates are shrunken toward the mean (the overall average fixed effects). These case-specific estimates are improved estimates compared to the single-level ordinary least squares estimates because information from the entire dataset is used (in other words, the empirical Bayes estimate is "borrowing strength" from all

available study evidence). For an introduction to empirical Bayes estimates, see Casella (1985). Instead of running three separate single-level analyses, one two-level hierarchical linear model can be run, providing both the effect size estimates across cases and case-specific estimates. The results of the case-specific estimates based on the empirical Bayes estimates are displayed in Table 2 and closely match the results of the single-level ordinary least squares analyses.

Multiple-baseline design – Alternating treatment design

In Alternating Treatment Designs (ATDs), two or more treatments (possibly following a baseline phase) are rapidly alternated (Barlow & Hayes, 1979; Barlow et al., 2009), or treatment sessions are alternated with no treatment sessions. Most of the ATDs are characterized by a baseline phase and two or more treatments, which are alternated during the treatment phase. In this scenario, the researcher is interested in the differential effect between the two treatment effects (i.e., the relative effectiveness of two or more interventions; Horner & Odom, 2014). Other ATDs are characterized by an alternation of two or more treatments, or with alternation of two or more treatments with baseline sessions. In this later scenario, a pure baseline comparison is not possible unless the alternation is proceeded or followed by a phase only including baseline measures (Zimmerman et al., 2019). If the baseline sessions are alternated with treatment comparisons from the beginning, it is unknown how the participants perform without being introduced to the treatment (which could be a confounding factor). In addition, multitreatment inference can occur as it can be the case that multiple treatments are effective because they are given in an alternated fashion (one treatment might strengthen the effectiveness of the other treatment and vice versa). Zimmerman et al. (2019) indicate that possible multitreatment interference can be detected with the inclusion of an initial baseline and visual analysis that compares the initial baseline level to the baseline observations that are part of the alternating sequence. Similarly, a phase for a specific treatment can be included so that the observations within the treatment phase can be compared to the treatment observations that are part of the alternating sequence.

To demonstrate a functional relation between the independent and dependent variables, the data from different treatments should not overlap. In addition, the ATD study should include at least four data points of comparison in each of the treatments and at least five repetitions of alternating sequence to meet the standards of *What Works Clearinghouse* (Horner & Odom, 2014; Kratochwill et al., 2010).

This combined SCED combines the unique strengths of ATDs with MBDs (i.e., external validity, making more generalized treatment effects). That is, the combination of ATDs with MBDs uses the rapid comparison of two or more conditions (ATDs) and the start of the intervention phase is staggered across participants (MBD). In this way, the combination of ATD + MBD allows identifying the treatment that has a larger effect with higher degrees of internal and external validity of measurements. Another possibility of the ATDs is that researchers may choose to continue only the treatments with the strongest effects in the final phases of the study (Kratochwill et al., 2010).

Statistical model Step 1: single-level analysis. Similar to the single-level (i.e., case-specific) analysis for the MBD + WRD, a case-by-case intervention effectiveness evaluation can be performed for MBD + ATD. More specifically, the following

research question is of interest: *"Is there a change in level for Treatment 1 and Treatment 2, respectively?"* The effect sizes of interest can be obtained by introducing dummy variables for each treatment. The dummy-coded variables, $Treatment_{mi}$s, indicate the treatment phase. For instance, $Treatment_{mi}$ equals one if the score belongs to treatment phase m on moment i, zero otherwise. If all the $Treatment_{mi}$s are zero, then the measurement occasion belongs to the baseline phase. For two treatments, the following regression equation can be used (using treatment indicators $Treatment_{1i}$ and $Treatment_{2i}$).

$$Y_i = \beta_0 + \beta_1 Treatment_{1i} + \beta_2 Treatment_{2i} + e_i \text{ with } e_i \sim N\left(0, \sigma_e^2\right) \quad (4)$$

β_0 indicates the baseline level, β_1 refers to the change in level between the baseline and Treatment 1 and β_2 refers to the change in level between the baseline and Treatment 2. The difference between β_1 and β_2 refers to the differential effect (e.g., "*Is one of the treatments relatively more effective*?"). Equation (4) can be extended by modeling linear or non-linear trends (Hembry et al., 2015; Moeyaert, Ugille, Ferron, Beretvas, et al., 2014), or adding more dummy variables in case more than two treatments are examined.

Statistical model Step 2: two-level analysis. This step is similar to Step 2 described for MBD + WRD design, where coefficients from the first level can be modeled as varying at the second level:

$$\text{Level 1}: \; Y_{ij} = \beta_{0j} + \beta_{1j} Treatment_{1i} + \beta_{2j} Treatment_{2i} + e_{ij} \text{ with } e_{ij} \sim N\left(0, \sigma_e^2\right) \quad (5)$$

Level 2:

$$\begin{cases} \beta_{0j} = \theta_{00} + u_{0j} \\ \beta_{1j} = \theta_{10} + u_{1j} \text{ with } \begin{bmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \end{bmatrix} \\ \beta_{2j} = \theta_{20} + u_{2j} \end{cases}$$
$$N \sim \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{u_0}^2 & \sigma_{u_0 u_1} & \sigma_{u_0 u_2} \\ \sigma_{u_1 u_0} & \sigma_{u_1}^2 & \sigma_{u_1 u_2} \\ \sigma_{u_2 u_0} & \sigma_{u_2 u_1} & \sigma_{u_2}^2 \end{bmatrix} \right) \quad (6)$$

This two-level analysis allows for making more generalized conclusions as overall average estimates across cases are obtained (the θs in Equation (6)). As noted before, case-specific estimates are available by requesting the empirical Bayes estimates. By estimating the variance/covariance matrix, the between-case variance in baseline level ($\sigma_{u_0}^2$) and treatment effect estimates ($\sigma_{u_1}^2$ and $\sigma_{u_2}^2$) can be obtained.

EMPIRICAL ILLUSTRATION

The study of Chung and Cannella-Malone (2010) will be used for the empirical demonstration. This study used an ATD that is characterized by a baseline phase followed by an alternating phase in which baseline and treatment sessions are alternated. In addition, the ATD is repeated across multiple independent participants, and the start of the randomization phase is staggered across the participants (MBD). The purpose of the Chung and Cannella-Malone study was to examine separate and combined effects of motivation operations of three participants with multiple disabilities in four pre-session conditions: (1) attention, (2) response blocking, (3) attention with response blocking, and (4) non-interaction. The dependent variable was stereotypic behavior, which was measured using the 10- partial interval recording. Inter-observer data were calculated for pre-session (39% of data) and treatment (40% of data) conditions, with the agreement reaching 98% and 99%.

The graphical display of the data can be found in Figure 3 (i.e., copied from the original study) and Figure 5 (i.e., recreated graph, using the retrieved data obtained with WebPlotdigitizer; Rohatgi, 2011).

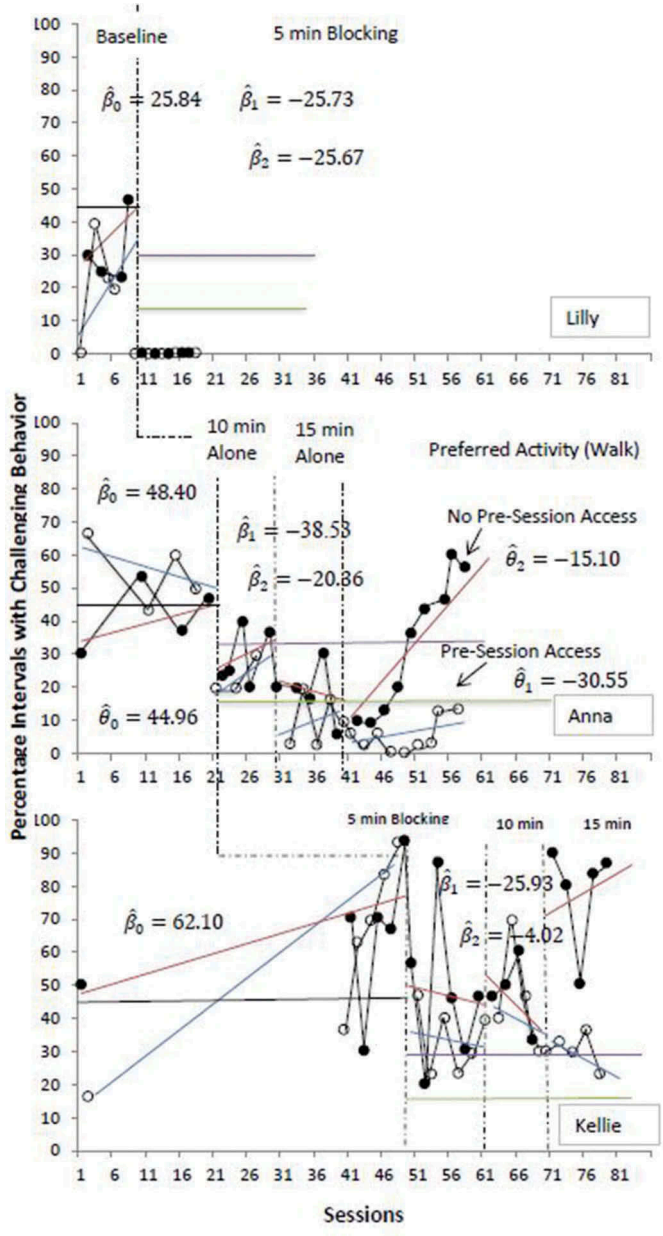


Figure 5. Estimated parameters for the single-level analysis and two-level analysis. The line during the baseline indicates the overall average baseline level estimate; the lines during the intervention indicate the estimated challenging behavior during the pre-session access intervention and the challenging behavior during the no pre-session access intervention.

For this empirical demonstration, we will analyze the problem behavior for the three participants of the study of Chung and Cannella-Malone (2010). During the treatment, participants did two tasks: Task A and Task B, which were individualized to the needs and skills of the participating students. Students did the tasks in two conditions as shown in Figure 3: (1) pre-session access condition that was identified in the functional analysis part of the study and (2) no pre-session access. Because of the individual needs in the Chung and Cannella-Malone (2010) study, the treatment phases are participant-specific. This is commonly the case using SCEDs as one of the strengths of this design is to adjust the treatment according to the participant's needs. As a consequence, the baseline versus treatment comparison for the three participants is not completely the same (i.e., Lilly: baseline – 5 min blocking; Anna: baseline – 10 min alone and Kellie: baseline – 5 min blocking). Therefore, strictly speaking, no experimental conclusions can be drawn from this combined design (Ledford and Gast, 2018). However, the treatment phases can be treated as subcategories of the same treatment and as a consequence it is still meaningful to investigate generalization of the effect across the three participants. In the original study, the data were visually analyzed, and the results were reported as percentages of intervals with problem behavior. Chung and Cannella-Malone (2010) reported that the intervention was successful for two out of the three participants, whose problem behaviors noticeably decreased. The results of the intervention for the third participant were contradictory (i.e., the intervention condition identified as successful in the previous experiment failed to decrease problem behaviors). Notwithstanding, the interventions were successful for only two out of the three participants, it is still worth estimating the size of the intervention effect to complement this finding. The coding of the design matrix for Case 1 (i.e., Lilly) in accordance with the mathematical model presented in Equation (4) can be found in Table 4. The SAS codes to run the analyses are available as a supplement to this article.

The output of the single-level analysis is presented in Table 5. From the case by case analysis, we can conclude that there is

Table 5. Results of ordinary least squares analysis and Empirical Bayes analysis per participant

Case	Parameter	Estimate (*SE*)	Estimate (*Standard error of prediction*)
Lilly	$\hat{\beta}_0$	25.84 (3.34)	26.69 (11.68)
	$\hat{\beta}_1$	−25.73 (5.39)	−30.46 (1.07)
	$\hat{\beta}_2$	−25.67 (5.39)	−21.60 (6.49)
Anna	$\hat{\beta}_0$	48.40 (4.52)	43.16 (11.58)
	$\hat{\beta}_1$	−38.53 (5.48)	−30.75 (1.07)
	$\hat{\beta}_2$	−20.36 (5.39)	−15.11 (6.01)
Kellie	$\hat{\beta}_0$	62.10 (5.81)	65.03 (11.56)
	$\hat{\beta}_1$	−25.93 (7.79)	−30.44 (1.07)
	$\hat{\beta}_2$	−4.02 (7.79)	−8.59 (6.03)

Table 6. Results of two-level analysis across participants

Parameter		**Estimate (*SE*)**	***t***	***p***
Fixed Effects				
Baseline level	$\hat{\theta}_0$	44.96 (11.67)	3.85	.057
Change in level Treatment 1	$\hat{\theta}_1$	−30.55 (4.10)	−7.44	.012
Change in level Treatment 2	$\hat{\theta}_2$	−15.10 (6.26)	−2.41	.152
Random Effects		**Estimate (*SE*)**	***z***	***p***
Baseline level	$\hat{\sigma}^2_{u_0}$	381.27 (399.56)	0.95	.17
Change in level Treatment 1	$\hat{\sigma}^2_{u_1}$	1.16 (43.95)	0.03	.489
Change in level Treatment 2	$\hat{\sigma}^2_{u_2}$	66.37 (122.25)	0.54	.293
Within-case variance	$\hat{\sigma}^2_e$	251.33 (36.64)	6.86	<.0001

a demonstration of treatment effectiveness for both interventions at two different points in time for Case 1 (i.e., Lilly) and for Case 2 (i.e., Anna) at the .05 significance level. When both pre-session and no pre-session access are introduced, we see a significant drop in problem behavior for Lilly [Case1 : $\hat{\beta}_1 = -25.73, t(15) = -4.77, p = .0002$ *and* $\hat{\beta}_2 = -25.67,\ t(15) = -4.76,\ p = .0003$], and Anna [Case2 : $\hat{\beta}_1 = -38.53,\ t(41) = -7.03, p < .0001$ *and* $\hat{\beta}_2 = -20.36, t(41) = -3.78, p = .0005$]. For Kelly (Case 3), there was only one demonstration of treatment effectiveness [$\hat{\beta}_1 = -25.93, t(39) = -3.33\ p = .0019$].

The two-level analysis was conducted to generalize treatment effectiveness beyond individual cases. Again, for didactic purposes, a small dataset with only three cases is used. In order to run a two-level analysis and obtain generalizable estimates, it is recommended to use a larger dataset. The results indicate that both the pre-session access and no pre-session access interventions succeeded in reducing the problem behaviors as negative estimates were obtained for the change in level between the baseline and Treatment 1 and the baseline and Treatment 2 [$\hat{\theta}_{10} = -30.55, t(61) = -7.44, p = .012; \hat{\theta}_{20} = -15.10, t(61) = -2.41, p = .152$]. However, only the estimate of the effect of Treatment 1 is statistically significant ($p < .05$). As can be seen in Table 6, the between-case variance in the treatment effects was large for Treatment 2 [$\hat{\sigma}^2_{u_2}$ = 66.37, Z = 0.54, p =.293], and the within-case residual variance is statistically significant [$\hat{\sigma}^2_e$ = 251.33, Z = 6.86, $p < .0001$].

The visual presentation of the single-level analysis and two-level analysis is given in Figure 5.

As mentioned earlier, an extra advantage of using the two-level model is that case-specific estimates are obtained in addition to the overall average estimates across cases. The results of the case-specific estimates based on the empirical Bayes estimates are displayed in Table 5 and closely resemble the results of the single-level analyses.

DISCUSSION

Previous research in the field of SCEDs solely focused on estimating intervention effectiveness using data from "single" SCEDs. This study expands on this and introduces an analysis technique suitable to estimate treatment effectiveness for more complex SCEDs, namely "combined SCEDs". This study is the first study to demonstrate how applied researchers can

use an extension of established methodology to come up with an effect size estimate appropriate for combined designs. The proposed technique is generic and not limited to combined designs. For instance, by excluding predictors in the two-level models, the technique can be used to quantify treatment effects across single SCEDs. Combined SCEDs are combinations of single SCEDs, and are frequently used as they are more internally and externally valid and can answer richer research questions. The two most popular combined designs are discussed in detail, namely the MBD + WRD and MBD + ATD. For these combined designs, we discuss (a) the mathematical models appropriate for the quantitative analysis, (b) the coding of the design matrix, (c) the statistical software to perform the analysis, (d) the interpretation of the output tables, and (e) the visual presentation of the obtained coefficients. We demonstrate the process using data from previously published studies. The purpose is to assist single-case researchers in drawing valid and reliable inferences regarding the treatment effectiveness for complex designs.

The single- and two-level hierarchical linear modeling (HLM) techniques are suggested. The two-level HLM is appropriate as both participant-specific and overall average study-specific estimates are obtained simultaneously (instead of running separate single-level analyses for each case), which leads to drawing more generalized inferences. Empirical Bayes estimates of the participant-specific treatment effects are more precisely estimated compared to the OLS (single-level) estimates, but they are biased toward the average effect. By ignoring the hierarchical structure of the data (i.e., measurements are nested within cases, and cases are nested within study), biased standard errors are obtained (the standard errors are too small due to ignoring the dependency), and, consequently, the analysis is prone to Type I errors. The two-level HLM provides regression-based effect size estimates and their standard errors. Therefore, they can be used afterward for meta-analytic purposes. A third level can be added to the model, and overall average treatment effectiveness can be estimated across studies. In addition, the variability in treatment effectiveness between studies can be explored. If a large amount of variability is identified, moderators can be added to the model. Another advantage of summarizing treatment effects across studies is the increased power to identify true treatment effects.

Limitations and future research directions

The HLM model introduced in this study is the most basic model, which ignores, for instance, data trend and autocorrelation, and is only appropriate for continuous outcomes. In addition, use of conventional HLM requires assumptions about multivariate normality that need to be met in order to make valid inferences (Raudenbush & Bryk, 2002). This was beyond the scope of this study as the focus was on the logic of modeling combined design SCEDs, which is already a complexity. However, use of the HLM is flexible, and other complexities can be introduced into the model. For instance, in case a researcher is studying a target behavior or skill in which a trend is expected, the introduced models can be extended by including a time indicator variable in the treatment phase. This results in two effect size estimators of interest: (1) change in level of the dependent variable when introducing the treatment and (2) the trend during the treatment phase. Two-level hierarchical linear modeling including a linear time trend is discussed in detail in Moeyaert, Ugille, Ferron, Beretvas, et al. (2014).

Another complexity relates particularly to the MBD + ATD design. In ATDs, the effectiveness of two (or more) treatments is compared with a common baseline phase, which introduces dependency. The model can be further extended by exploring options to model this dependency (by, for instance, estimating the covariance or using a more complex estimation technique if more cases within a study are included, specifically robust variance estimation; Hedges et al., 2010). Last, when using HLM, caution needs to be exercised when interpreting the between-case variance estimates as severely biased estimates can be obtained (Moeyaert et al., 2013). The limitations discussed here are not specific to HLM of combined SCEDs, but for using HLM in general as an analysis technique for the quantitative integration of SCED data.

In addition, the results of the two studies discussed in this article should be interpreted with caution because in both of them there was a lack of experimental control. In Seybert et al. (1996), the withdrawal and reversal design embedded in the combined design did not meet the basic replication standards for one of the participants. In addition, there was a non-effect for the withdrawal of the treatment for that same participant. As a consequence, to meet the WWC design standards to demonstrate experimental control, there is an additional basic replication needed for one of the participants of the Seybert et al. (1996) study. Similarly, in Chung and Cannella-Malone (2010), the treatment to reduce problem behaviors was effective for two out of three participants. In addition, the effectiveness of the treatment was investigated across slightly different treatment phases. In order to meet the WWC design standards, the treatment phases across the participants should be identical and there should be three demonstrations of the effectiveness of the treatment at three different points in time. Effect size estimation for these combined designs is still informative as it quantifies the magnitude of treatment effect. This quantification provides an overall summary of the study findings (and variability between participants in treatment effectiveness) and can be used for meta-analysis purposes afterward. However, we encourage applied SCED researchers designing combined SCEDs that meet the WWC design standards for experimental control. In order to demonstrate our methodology, we were limited to published combined designs. The examples included are typical for the field and are solely used to demonstrate the analysis technique.

In terms of future research directions, the suggested models can be extended by adding case characteristics (gender, age, race, etc.) to investigate their moderating effect on the treatment effectiveness. However, recent research related to power indicates that at least 12 cases are needed, or 7 cases in combination with at least 40 measurement occasions, to be able to include case characteristics in the analyses (Moeyaert et al., 2017). This, of course, depends on the particular predictors and the value of the true treatment effect. Simulation studies can be performed in order to investigate the power for a particular set of design conditions. Again, this is beyond the scope of this paper. Other ways of coding the design matrix are also possible depending on the specific research questions and structure of the data being analyzed.

To further enhance the internal validity, single-case researchers might consider introducing randomization when developing the combined SCED design. As discussed in depth by J. R. Ledford et al. (2018), several forms of randomization can be incorporated in the design. First, the start and the retrieval of the intervention can be randomized. In this scenario, it is recommended that the randomization does not start until baseline stability is established. Second, the order of the conditions

can be randomized, which is typically done in ATDs. Unrestricted randomization is not recommended to avoid conditions not representing ATDs (i.e., all baseline conditions could be chosen first) or to avoid that a certain randomized pattern is consistently chosen (i.e., treatment 1 is always administered after treatment 2). A third randomization form is the random assignment of participants to intervention start points. This is relevant for multiple-baseline designs across participants. Incorporating randomization in the design allows for use of randomization tests to make conclusions related to treatment effectiveness. The advantage of such tests is that the sampling distribution is built based upon the randomization patterns and as a consequence, no parametric assumptions are made and needed (for more details about randomization, see J. M. Ferron & Levin, 2014; Heyvaert et al., 2017). Inclusion of randomization has the potential to reduce the risk of biased effect size estimates.

In order to increase the external validity of treatment effectiveness and contribute to evidence-based decisions in research, practice and policy, multiple SCED studies can be summarized. Previous research demonstrates how the multilevel meta-analytic framework can be used to combine single SCEDs (Moeyaert, 2018; Moeyaert, Ugille, Ferron, Onghena, et al., 2014). Therefore, future research is needed to demonstrate how pure and combined SCEDs can be combined using the multilevel meta-analytic approach. Similarly, a following-up study can be conducted to evaluate the consequences of ignoring the complex nature of combined designs.

CONCLUSIONS

This study is the first study introducing and demonstrating a promising methodological framework for effect size estimation for combined SCEDs. The two-level hierarchical model is recommended as it has the possibility to include variables to account for the combined design complexity. In this study, the logic of modeling the combined SCED study is introduced, empirical illustrations are given, analysis output is discussed and SAS code is supplemented. Single-case researchers are given the tools (and are encouraged) to modify and/or further extend the models. The proposed method of coding and estimating effect sizes for combined SCEDs can be a useful technique to inform researchers and practitioners about the effectiveness of interventions.

DISCLOSURE STATEMENT

No potential conflict of interest was reported by the author(s).

Funding

This research was supported by the Institute of Education Sciences, U.S. Department of Education, through grants [R305D150007 and R305D190022]. 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

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. doi: 10.3758/s13428-012-0231-z

Barlow, D. H., & Hayes, S. C. (1979). Alternating treatments design: One strategy for comparing the effects of two treatments in a single subject. *Journal of Applied Behavior Analysis*, *12*(2), 199–210. doi: 10.1901/jaba.1979.12-199

Barlow, D. H., Nock, M., & Hersen, M. (2009). *Single case experimental designs : Strategies for studying behavior for change*. Pearson/Allyn and Bacon.

Casella, G. (1985). An introduction to empirical Bayes analysis. *The American Statistician*, *39*(2), 83–87. doi: 10.2307/2682801

Chung, Y., & Cannella-Malone, H. I. (2010). The effects of profession manipulations on automatically maintained challenging behavior and task responding. *Behavior Modification*, *34*(6), 479–502. doi: 10.1177/0145445510378380

Ferron, J., & Jones, P. K. (2006). Tests for the visual analysis of response-guided multiple-baseline data. *The Journal of Experimental Education*, *75*(1), 66–81. doi: 10.3200/JEXE.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: Methodological and statistical advances* (pp. 153–183). American Psychological Association.

Gast, D. L., Lloyd, B. P., & Ledford, J. R. (2018). Multiple baseline and multiple probe designs. In J. R. Ledford & D. L. Gast (Eds.), *Single case research methodology* (3rd ed., pp. 239–281). Routledge.

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta- regression with dependent effect size estimates. *Research Synthesis Methods*, *1*(1), 39–65.

Hembry, I., Bunuan, R., Beretvas, S. N., Ferron, J., & Van den Noortgate, W. (2015). Estimation of a nonlinear intervention phase trajectory for multiple-baseline design data. *The Journal of Experimental Education*, *83*(4), 514–546. doi: 10.1080/00220973.2014.907231

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), 493–590. doi: 10.1111/jar.12094

Heyvaert, M., Moeyaert, M., Verkempynck, P., Van den Noortgate, W., Vervloet, M., Ugille, M., & Onghena, P. (2017). Testing the intervention effect in single-case experiments: A Monte Carlo simulation study. *The Journal of Experimental Education*, *85*(2), 175–196. doi: 10.1080/00220973.2015.1123667

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

Jason, L. A., & Frasure, S. (1979). *Increasing peer-tutoring behaviors in the third grade classroom* [Paper presentation]. *Annual Convention of the American Psychological Association*, New York.

Kelley, M. E., Lerman, D. C., & Van Camp, C. M. (2002). The effects of competing reinforcement schedules on the acquisition of functional communication. *Journal of Applied Behavior Analysis*, *35*(1), 59–63. doi: 10.1901/jaba.2002.35-59

Kennedy, C. (2005). *Single-case designs for educational research* (Vol. 1). Pearson/A & B.

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. doi: 10.1007/s10803-009-0931-0

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*. What Works Clearinghouse. https://ies.ed.gov/ncee/wwc/Docs/ReferenceResources/wwc_scd.pdf

Kratochwill, T. R., & Levin, J. R. (2014). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Statistical and methodological advances* (pp. 53–91). American Psychological Association.

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: Statistical and methodological advances* (pp. 91–125). American Psychological Association.

Ledford, J., & Gast, D. L. (2018). Combination and other designs. In J. R. Ledford & D. L. Gast (Eds.), *Single case research methodology* (3rd ed., pp. 239–281). Routledge.

Ledford, J., Lane, J., & Severini, K. (2018). Systematic use of visual analysis for assessing outcomes in single case design studies. *Brain Impairment*, *19*(1), 4–17. doi: 10.1017/BrImp.2017.16

Ledford, J. R., Lane, J. D., & Tate, R. (2018). Evaluating quality and rigor in single case research. In J. R. Ledford & D. L. Gast (Eds.), *Single case research methodology* (3rd ed., pp. 365–392). Routledge.

Lenz, A. S. (2013). Calculating effect size in single-case research: A comparison of nonoverlap methods. *Measurement and Evaluation in Counseling and Development*, *46*(1), 64–73. doi: 10.1177/0748175612456401

Lindberg, J. S., Iwata, B. A., & Kahang, S. W. (1999). On the relation between object manipulation and stereotypic self-injurious behavior. *Journal of Applied Behavior Analysis*, *32*(1), 51–62. doi: 10.1901/jaba.1999.32-51

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. doi: 10.1016/j.jsp.2011.03.004

Manolov, R., & Solanas, A. (2013). A comparison of mean phase difference and generalized least squares for analyzing single-case data. *Journal of School Psychology*, *51*(2), 201–215. doi: 10.1016/j.jsp.2012.12.005

Matson, J. L., & Keyes, J. B. (1990). A comparison of DRO to movement suppression time-out and DRO with two self-injurious and aggressive mentally retarded adults. *Research in Developmental Disabilities*, *11*(1), 111–120. doi: 10.1016/0891-4222(90)90008-V

Moeyaert, M., Ugille, M., Ferron, J., Beretvas, S., & Van den Noortgate, W. (2013). Three-level analysis of single-case experimental data: Empirical validation. *Journal of Experimental Education*, *82*(1), 1–21. doi: 10.1080/00220973.2012.745470

Moeyaert, M., Ugille, M., Ferron, J., Beretvas, S., & Van den Noortgate, W. (2014). 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. doi: 10.1177/0145445514535243

Moeyaert, M., Ugille, M., Ferron, J., Onghena, P., Heyvaert, M., & Van den Noortgate, W. (2014). Estimating intervention effects across different types of single-subject experimental designs: Empirical illustration. *School Psychology Quarterly*, *25*(1), 191–211. doi: 10.1037/spq0000068

Moeyaert, M., Ferron, J., 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. doi: 10.1016/j.jsp.2013.11.003

Moeyaert, M., Maggin, D. M., & Verkuilen, J. (2016). Reliability and validity of extracting data from image files in contexts of single-case experimental design studies. *Behavior Modification*, *40*(6), 874–900. doi: 10.1177/0145445516645763

Moeyaert, M., Akhmedjanova, D., & Bogin, D. (2017). *The power to test moderator effects in multilevel modeling of single-case data* [Manuscript in preparation].

Moeyaert, M., Zimmerman, K., & Ledford, J. (2018). Analysis and meta-analysis of single-case experimental data. In J. Ledford & D. Gast (Eds.), *Single-case methodology: applications in special education and behavioral sciences*. New York: Routledge.

Moeyaert, M. (2019). Quantitative synthesis of research evidence: Multilevel meta-analysis. *Behavioral Disorders*, *44*(4), 241–256. doi: 10.1177/0198742918806926

Moeyaert, M., Klingbeil, D., Rodabaugh, E., & Turan, M. (2019). Multilevel meta-analysis of peer-tutoring interventions to increase academic performance and social interactions for people with special needs. *Remedial and Special Education*. doi: 10.1177/0741932519855079

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

Parker, R. I., Vannest, K. J., & Davis, J. L. (2014). A simple method to control positive baseline trend within data nonoverlap. *The Journal of Special Education*, *48*(2), 79–91. doi: 10.1177/0022466912456430

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage.

Rohatgi, A. (2011). *WebPlotDigitizer*. https://automeris.io/WebPlotDigitizer

Schlosser, R. W., Lee, D. L., & Wendt, O. (2008). Application of the percentage of non-overlapping data in systematic reviews and meta-analyses: A systematic review of reporting characteristics. *Evidence-based Communication Assessment and Intervention*, *2*(3), 163–187. doi: 10.1080/17489530802505412

Scruggs, T. E., Mastropieri, M. A., & Casto. (1987). The quantitative synthesis of single subject research: Methodology and validation. *Remedial & Special Education*, *8*(2), 24–33. doi: 10.1177/074193258700800206

Seybert, S., Dunlap, G., & Ferro, J. (1996). The effects of choice-making on the problem behaviors of high-school students with intellectual disabilities. *Journal of Behavior Education*, *6*(1), 49–65. doi: 10.1007/BF02110477

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. doi: 10.1080/17489530802581603

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. doi: 10.3758/s13428-011-0111-y

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. doi: 10.1037/a0032964

Shadish, W. R., Rindskopf, D. M., Hedges, L. V., & Sullivan, K. J. (2013). Bayesian estimates of autocorrelations in single-case designs. *Behavior Research Methods*, *45*(3), 813–821. doi: 10.3758/s13428-012-0282-1

Shadish, W. R., Zuur, A. F., & Sullivan, K. J. (2014). Using generalized additive (mixed) models to analyze

single case designs. *Journal of School Psychology*, *52*(2), 149–178. doi: 10.1016/j.jsp.2013.11.004

Shogren, K. A., Faggella-Luby, M. N., Bae, S. J., & Wehmeyer, M. L. (2004). The effect of choice-making as an intervention for problem behavior: A meta-analysis. *Journal of Positive Behavior Interventions*, *6*(4), 228–237. doi: 10.1177/10983007040060040401

Swaminathan, H., Horner, R. H., Sugai, G., Smolkowski, K., Hedges, L., & Spaulding, S. A. (2010). *Application of generalized least squares regression to measure effect size in single-case research: A technical report*. *Unpublished technical report*, Institute for Education Sciences.

Trottier, N., Kamp, L., & Mirenda, P. (2011). Effects of peer-mediated instruction to teach use of speech-generating devices to students with autism in social game routines. *Augmentative and Alternative Communication*, *27*(1), 26–39. doi: 10.3109/07434618.2010.546810

Wang, S. Y., Parrila, R., & Cui, Y. (2013). Meta-analysis of social skills interventions of single- case research for individuals with autism spectrum disorders: Results from three-level HLM. *Journal of Autism and Developmental Disorders*, *43*(7), 1701–1716. doi: 10.1007/s10803-012-1726-2

Wolery, M., Busick, M., Reichow, B., & Barton, E. E. (2010). Comparison of overlap methods for quantitatively synthesizing single-subject data. *The Journal of Special Education*, *44*(1), 18–28. doi: 10.1177/0022466908328009

Zimmerman, K. N., Ledford, J. R., & Severini, K. E. (2019). Brief Report: The effects of a weighted blanket on engagement for a student with ASD. *Focus on Autism and Other Developmental Disabilities*, *34*(1), 15–19. doi: 10.1177/1088357618794911