Meta-analysis of single-case research via multilevel models: Review of concepts and empirical evidence

Mariola Moeyaert[1], Rumen Manolov[2] and Emily Rodabaugh[1]

[1] Department of Educational Psychology and Methodology, State University of New York, NY

[2] Department of Social Psychology and Quantitative Psychology, University of Barcelona, Spain

Running head: HLM for SCED meta-analysis

Contact author

Correspondence concerning this article should be addressed to Mariola Moeyaert, Department Educational and Counseling Psychology, State University of New York, NY, 1400 Washington Ave., 12222, Albany, US. Phone number: +1 (518) 618-6056. E-mail: mmoeyaert@albany.edu

Abstract

Multilevel models were developed to analyze hierarchical structured data with units at a lower level nested within higher-level units. Single-case experimental design (SCED) data are collected from multiple cases within a study and allow researchers to investigate intervention effects at the individual level and also to investigate how these individual intervention effects change over time. Given the increased interest in establishing an evidence base for interventions, SCED data from multiple studies can be synthesized; therefore a multilevel meta-analysis is appropriate. Although using multilevel models to meta-analyze SCED studies is promising, their actual implementation is hampered by being potentially excessively technical. Therefore, this article provides an accessible description and overview of the potentials of multilevel meta-analysis to combine SCED data. Moreover, a summary of the evidence on the performance of multilevel models for meta-analysis is provided, which is useful given that such evidence is currently scattered over multiple technical methodological articles.

Key words: Single-case experimental design, HLM; multilevel; meta-analysis

Meta-analysis of single-case research via multilevel models: Review of concepts and empirical evidence

In the context of the need to establish the evidence basis of interventions, single-case experimental designs (SCEDs) have been considered a valid way of providing such evidence due to their internal validity (Horner et al., 2005; Howick et al., 2011; Kratochwill & Levin, 2010). In relation to internal validity, several attempts to observe the effect of the intervention are required (Kratochwill et al., 2010), which is achieved via direct replication (Kennedy, 2005) and usually by means of a multiple-baseline design (Shadish & Sullivan, 2011; Smith, 2012). In terms of generalizability of the results, Kennedy (2005) argues that it is advisable to distinguish between external validity (informing about how representative the results are of a greater population) and systematic replication (informing about how the intervention works in different conditions). In relation to external validity, there is a wide support for the need to synthesize the information obtained from several SCED studies (e.g., Beeson & Robey, 2006; Jenson, Clark, Kircher, & Kristjansson, 2007; Maggin, Lane, & Pustejovsky, 2017; Schlosser, 2005). As a result, more than 100 meta-analyses of SCED studies have already been performed in several disciplines (see the reviews by Maggin, O'Keefe, & Johnson, 2011 and Jamshidi et al., 2017). In this context, HLMs are suggested and are approproate to quantitatively integrate results across SCED studies. HLMs are useful for obtaining both overall summary statistics and a quantification of the variability in effectiveness of a treatment across studies as a means for assessing the generalizability of findings. Additionally, HLMs easily incorporate moderator analysis, which could be useful to explain variability in the effectiveness of a treatment between SCED studies. Finally, although it is not the focus of the current study, HLMs are consistent with the importance of replication within a SCED study, due to three reasons: (a) HLMs account for

the nested structure of the data (measurements within participants within studies); (b) it is possible to quantify the amount of variability across replications within a study; and (c) it is possible to use HLMs, as a visual tool, as part of the assessment of consistency of the effect across replications (Manolov, 2017) .

The aim of the current study is to provide a broad overview of HLMs in such a form that is accessible to applied researchers. In order to achieve this aim, verbal descriptions are offered of the main features, strengths, and limitations of HLMs, whereas the reader interested in more technical description of HLMs or in articles describing their application to several datasets can consult the references provided in the Appendix. This overview includes four key components (included in the following sections), none of which are available elsewhere in the literature: (a) step-by-step summary of this rather complex data analytical technique in a non-technical, comprehensive and accessible way, in absence of formulas and with the main technical details explained in plain language; (b) an illustration of how to use a multilevel analysis and a modified Brinley plot (Blampied, 2017) together for assessing the degree to which an intervention tested in several studies is effective; (c) a summary of previously published empirical evidence on the performance of multilevel models, so the researchers can easily identify in which conditions the multilevel modeling is applicable, valid and reliable; and (d) recommendations for conducting research (in addition to the methodological quality indicators already available, e.g., Horner et al., 2005; Reichow, Volkmar, & Cicchetti, 2008; Tate et al., 2013), for data analysis, and for the roles of applied researchers and methodologists/statisticians when performing a multilevel meta-analysis of SCED data.

Given that, in this study, the focus of the application of HLMs is meta-analysis, first its main features and components are briefly introduced. More details on meta-analysis can be obtained

from some excellent textbooks (e.g., Cooper, Hedges, & Valentine, 2009; Lipsey & Wilson, 2001) and some articles specifically dedicated to the (meta-)analysis of SCEDs (e.g., Beretvas & Chung, 2008: Onghena et al., 2018).

A Brief Review of Meta-Analysis

Procedure

A meta-analysis is guided by a research question (e.g., whether an intervention for a specific problem is effective and how effective) and by search parameters, such as keywords, databases, and eligibility criteria for the studies to be included in the quantitative integration. A meta-analysis, in general, answers the question about how large the intervention effect is. It is common to obtain a weighted average on the basis of a random-effects model, according to which the study effects vary across studies because of sampling variation (as in the fixed-effect model), but also because of real differences between the studies (e.g., due to different sampling methods, experimental manipulations, outcome measures). A HLM that uses a null model (i.e., only with the intercept but with no predictors) is equivalent to a random-effects model. There is no clear evidence that either the multilevel approach or traditional random-effects approaches are superior (Van den Noortgate & Onghena, 2003b).

Handling Dependencies

When performing meta-analyses, it is important to deal with any potential dependencies between the effect sizes being integrated. One possible situation is that, in each study, there are several different outcomes of interest (i.e., different response variables). For such a situation, a separate univariate meta-analysis can be performed for each measure. Another possible situation is that there are several outcomes representing several replications for the same response

variable. For such a situation, the average of the outcomes can be obtained and this would lead to having one effect size per study, but this is not optimal (Moeyaert, Rindskopf, Onghena, & Van den Noortgate, 2017). For both situations, an alternative approach is to use a multivariate model (Kalaian & Raudenbush, 1996) or a multilevel model with an intermediate level (Van den Noortgate, López-López, Marín-Martínez, & Sánchez-Meca, 2013). In the latter case, individual measurements (level-1) would be nested into outcomes (level-2), which would be nested into studies (level-3). Finally, robust variance estimation can be used for handling dependencies (Hedges, Tipton, & Johnson, 2010).

Moderator Analysis

The aim of moderator analysis is to identify study characteristics that explain the variability in the effect sizes. The analysis is performed as a weighted regression using study characteristics as predictors of the effect sizes; the weights are the sample sizes per study. The moderators are preferably selected on a theoretical basis in order to avoid testing too many predictor variables, which could affect statistical power.

Publication Bias

A sample of studies that is not representative of all studies conducted on the topic can affect the validity of the conclusion of meta-analysis. Several procedures have been suggested for dealing with this issue: (a) the fail-safe *N* (Rosenthal, 1979) represents an assessment of the robustness of the evidence against publication bias; (b) the funnel plot and the Egger test (Egger, Smith, Schneider, & Minder, 1997) for funnel plot asymmetry can be used to detect whether studies with small effect sizes are missing; and (c) the trim-and-fill method (Duval, 2005; Duval

& Tweedie, 2000a, 2000b) re-estimates the overall effect size in the event that potentially missing studies were present in the meta-analysis.

Multilevel Models and Their Application to SCED Meta-Analysis

Concepts

A HLM is called hierarchical because it refers to situations in which the data follow a nested structure: there are higher levels and lower levels, as will be explained in detail in the next section. The term "multilevel models" differs slightly from HLM. In a multilevel model, the relation between the predictor and the outcome variable is not assumed to be linear. For instance, it could be quadratic and exponential (Hembry, Bunuan, Beretvas, Ferron, & Van den Noortgate, 2015). HLMs or multilevel models in general are sometimes referred to as "mixed-effects models" or simply "mixed models" because they entail fixed effects (e.g., estimating an average effect across participants) and random effects (i.e., allowing to estimate variability in effect size effectiveness across participants and studies). This variability can be accounted for or explained by including moderator variables (i.e., covariates, predictors) at different levels. The reader interested in more advantages and possibilities of multilevel models is referred to Raudenbush and Bryk (2002).

Data with a Nested Structure

HLMs are especially applicable when the data is characterizes by a nested structure: for instance, repeated measurements (level-1 units) nested within participants (level-2 units). Accounting for dependencies arising from this nested structure leads to unbiased standard error

estimates of the coefficients of interest and thus to obtaining correct *p* values. Nevertheless, in the event that practically all of the variability observed in the response variable is between level-1 units (e.g., between the measurements of a participant students) and not between the level-2 units (e.g., between the participants), this would imply that a HLM is not justified (Gage & Lewis, 2014). That is, if there is no actual clustering (i.e., similarities within the level-2 units and differences between the level-2 units), the nested structure of the data is not quantitatively clear[1].

When HLM is used for meta-analysis of SCED studies, several multilevel modeling structures can occur. The most common multilevel structures are represented in Table 1. The researchers carrying out the meta-analyses need to prepare their data files according to the actual data structure. In the following paragraphs, some specific issues that need to be kept in mind are discussed.

A first comment related to preparing the data refers to deciding how to deal with data from ABAB reversal designs. We concur with Pustejovsky and Ferron (2017) who state that it is better to use all data and perform the logical comparisons A_1-B_1 and A_2-B_2 (A_1 and A_2 refer to Baseline 1 and Baseline 2 respectively whereas B_1 and B_2 refer to the Treatment 1 and Treatment 2 respectively) . In a multilevel meta-analysis, the AB-comparisons would represent a separate level (see Structure 2 from Table 1).

A second comment related to the importance of preparing the data refers to the need to standardize the raw scores if the outcome scale differs across studies. The proposal by Van Den Noortgate and Onghena (2008) for making the scores comparable is to divide them by the within-case residual standard deviation obtained from either an ordinary least squares regression

[1] For more information on data aggregation in multilevel analyses, the reader is referred to Dixon and Cunningham (2006).

(if time trend is not modeled) or from a piecewise regression if the time trends in the baseline and treatment phases are separately modeled. That is, an initial regression is carried out for each participant using the dummy phase variable as predictor, and afterwards, the individual scores are divided by the standard deviation of the (within-subject) residuals, also known as the root mean square error. It should be noted that standardizing would lead to losing information about the initial baseline level, which is potentially relevant for interpreting the magnitude of the effect in substantive terms. However, this is the case for any type of standardization.

A third comment related to the importance of preparing the data refers to Structure 4 from Table 1. In order to perform a HLM meta-analysis for such a situation, the effect sizes should have a known variance that can be estimated even in absence of raw data. Van Den Noortgate and Onghena (2008) provide the formula for estimating the variance of the standardized mean difference or a regression beta coefficient when the predictor is a phase dummy. Using effect sizes as level-1 units can be more efficient (i.e., including fewer [fixed and random] coefficients).

INSERT TABLE 1 ABOUT HERE

Information Obtained at Each Level

HLMs are efficient because the initial analysis provides regression coefficients for the highest level only (i.e., on average for all studies included in a meta-analysis or on average for all participants within a study), but there is a further analysis possible for the lower level estimates (i.e., case-specific and study-specific effect sizes). If a researcher is interested in the individual regression coefficients at lower levels (i.e., for each participant within a study and/or for each study included in the meta-analysis), empirical Bayes techniques can be used afterwards (Ferron,

Farmer, & Owens, 2010). These individual estimates are shrunken towards (i.e., closer to) the overall average; there is more shrinking for cases in which fewer measurements are obtained. The aim is to gain precision even for the participants for whom the series are shorter or for studies with a smaller amount of participants (Dedrick et al., 2009; Onghena, Michiels, Jamshidi, Moeyaert, & Van den Noortgate, 2018).

Predictors in the SCED Context

The commonly used predictor variables at level-1 are session number (for modeling general time trend), a dummy coded variable (0-1) representing the phase, and an interaction term representing the relation between the dummy variable and general time trend. The first predictor, general time trend refers to the slope of the baseline data expected to continue into the intervention phase. In the dummy predictor, 0 represents the data from the baseline phase and 1 the data from the intervention phase. Finally, the interaction term can be understood as a way to model change in slope between baseline and intervention phase (also called the effect of the intervention on time trend), given that it allows for the trend in the intervention phase to be different from the trend in the baseline phase. All these predictors are called "time-variant" (Shaw & Liang, 2012), as they change with time.

If the interaction term is not in the model, the dummy predictor represents the average change in level between the baseline phase and the treatment phase. If only the dummy predictor is in the model (e.g., when there is no general trend and no change in slope), the model is equivalent to the one used in BC-SMD (Shadish, Hedges, Pustejovksy, 2014), but the estimation procedure is different (maximum likelihood or restricted maximum likelihood for HLM and moment estimation for BC-SMD). Moreover, the BC-SMD provides an overall estimate of the change in

level expressed in standard deviations, whereas HLMs provide raw estimates of the change in level (i.e., in the same measurement units as the dependent variable). This is beneficial as applied researchers can then use these raw estimates to make inferences in the same units as the dependent variable.

In the context of a meta-analysis of SCED data, further predictors at level-2 (characteristics of the participant and the intervention) and at level-3 (characteristics of the study) can be introduced. These predictors are part of the moderator analysis.

Centering of Predictors

If general time trend is not modeled (i.e., the predictor representing session number is not present in the model), the intercept represents the average baseline level because for the entire baseline, the value of the dummy predictor is equal to 0. If general time trend and the effect of the intervention on the time trend (i.e., change in slope) are modeled, centering becomes especially important for interpreting the intercept and the effect associated with the dummy variable. One option would be to center only the interaction term (representing change in slope between the baseline and the treatment), but not the general time variable. The notation presented and explained by Huitema and McKean (2000; their Table 1) refers to such a situation. The coefficient associated with the interaction term refers to the change in slope starting from the first intervention phase measurement occasion. The coefficient associated with the dummy predictor refers to the immediate change at the first intervention phase measurement occasion and, in case there is no general time trend and no change in slope, it is equivalent to the mean phase difference. Another option is to center both general time and the interaction term (Moeyaert, Ugille, Ferron, Beretvas, & Van den Noortgate, 2014). Such centering usually entails

setting the initial baseline measurement occasion to 0, which makes the intercept represent the expected value at the beginning of the baseline. The interpretation of the dummy predictor and the interaction term is maintained. Further information on centering in the context of single-case data can be found in Chapter 6 of Raudenbush and Bryk (2002) and for other kinds of data in Dedrick et al. (2009; more accessible) and Kreft, de Leeuw, and Aiken (1995; more technical).

Modeling Options in Relation to Level-1 Predictors

The choice of the data aspects to model at the first level (when working with raw or standardized raw data but not when working with effect sizes directly) can be made on a theoretical basis. For instance, if a spontaneous recovery is expected, general time trend should be modelled; if the effect is expected to be progressive, change in slope should be modelled. Another option is to guide the choice of what data aspects to model using a visual inspection (Baek, Petit-Bois, Van den Noortgate, Beretvas, & Ferron, 2016; Moeyaert et al., 2014). However, the second option could potentially bias the results and, thus, should not become standard practice. Nevertheless, visual analysis (e.g. using the *multiSCED* tool from http://52.14.146.253/MultiSCED/) could be useful in order to validate the results obtained by the quantitative analysis: that is, in order to assess whether the quantifications are meaningful for the data pattern actually obtained (Parker, Cryer, & Byrns, 2006). In case a second model is selected and run, it should be stated clearly that it is chosen not a priori, but on the basis of the data at hand. Later in the text, sensitivity analysis is discussed, which is related to model selection.

Descriptive Results

When the term “estimation” is used, it refers to the descriptive values obtained for quantifying the average (fixed) treatment effects and the variances (random effects) around these

averages. These descriptive values (averages and the square root of the variance, which is the standard deviation) are expressed in the same measurement units as the response variable. The estimates can be obtained via Full Maximum Likelihood (FML) estimation (in which the fixed and random effects are estimated jointly) or via Restricted Maximum Likelihood (REML) estimation (in which the estimates of the random effects are obtained first). (Actually, HLM could be understood an extension of piecewise regression [Center, Skiba, & Casey, 1985-1986], but the latter uses ordinary least squares estimation.).

REML estimation is less biased and even optimal for equally-sized level-2 units (Hox, 2010), but does not allow for comparing models that differ in their fixed effects. Specifically, for the purpose of synthesizing SCED studies, REML is recommended (Ferron, Bell, Hess, Rendina-Gobioff, & Hibbard, 2009; Owens & Ferron, 2012) because it performs better (i.e., unbiased treatment effects and less biased variance estimates are obtained) when fewer units are available. REML is also used in a new definition of a standardized mean difference for SCED data on the basis of multilevel models (Pustejovsky, Hedges, & Shadish, 2014). Other options include bootstrap estimation, especially useful for studies with small sample and/or non-normal data (Van den Noortgate & Onghena, 2005), and Bayesian estimation (Moeyaert et al., 2017).

Inferential Results

Apart from the descriptive values (i.e., the estimates), it is possible to assess whether the values are statistically significantly different from zero. For the fixed effects, statistical significance is tested usually via the Student's t-test in which the estimate is divided by its standard error and referred to a student's t-distribution. Confidence intervals can be constructed for the fixed effects and are calculated based on the point estimates, the standard error estimates,

and one of several possible methods for estimating the degrees of freedom. For instance, in SCED context, the Satterthwaite or the Kenward-Roger estimated degrees of freedom have been shown to perform best (Ferron et al., 2009; Owens & Ferron, 2012). For small samples, it is preferred to construct confidence intervals rather than to obtain *p*-values via the Student's t-distribution. Standard statistical software packages (e.g., SPSS and SAS) use the Wald test to evaluate statistically significance of the random components estimates. Using this test, the estimate is divided be the standard error estimate and compared to the Z-(standard normal) distribution. However, this leads to invalid inferences as the sampling distribution of the random components is not normally distributed (negative values are not possible). This is the reason why the software package R does not display the p-values for the random components estimates. Instead, the statistical significance of the random effects can be assessed via a chi-square test on the deviances of the model with and without the random effect; i.e., it is based on comparing models (Hox, 2010).

Meta-analysis Using HLM

The results of interest in a multilevel meta-analysis is usually the following: (a) the average intervention effect estimate (across participants and across studies) – a descriptive estimate (i.e., a fixed effect) and a *p* value can be obtained; (b) the amount of variability in treatment effect estimate across studies (i.e., random effect, between-study variance); and (c) the amount of variability in treatment effect estimate across participants (i.e., random effect, between-case variance). It can also be informative to explain variability in treatment effect estimates between cases and/or between studies by modeling moderators (i.e., explaining why the intervention is larger or smaller in some studies and/or for some participants. This is discussed in the Moderator Analysis section.

Model Comparison

Comparing models is a common part of the model building procedure that usually takes place when a HLM is used (See Chapter 4 in Hox [2010] and Chapter 6 in Snijders and Bosker [1999] for a more theoretical information and Gage and Lewis [2014] and Wang, Parrila, and Cui [2013] for two examples.) When the models are nested (i.e., one is a simpler version of the other, that is, it contains fewer parameters), a chi-square test (also known as a likelihood ratio test) can be used to test whether the difference between the deviances (quantifications of unexplained variability) of the models are statistically significant. Such a test is useful for assessing the importance of random effects. Moreover, if FML is used, a chi-square test can also be used to assess the statistical significance of a fixed effect (i.e., whether it is necessary to include an additional predictor, such as a moderator).

Regardless of whether the models are not nested or not, there are other ways of comparing models, such as the Akaike Information Criterion (AIC) and Schwartz's Bayesian Information Criterion (BIC). Both are based on the deviances of the compared models and both penalize for the number of parameters, i.e., favor models that are more parsimonious. The difference is that BIC includes an additional penalty for more complex models. Both AIC and BIC are applicable for comparing models that differ in the fixed or random part when FML estimation is used. These criteria are also applicable for comparing models that differ in the random part when REML is used.

Moderator Analysis

Predictors can be included at different levels for accounting for the between-study variance in the effect sizes. Moreover, HLM offers an additional advantage in that it is possible to add a

moderator variable only to the random part, excluding it from the fixed part (e.g., for artifacts suspected to affect the variability of the effect sizes but not their average value). When the predictor variables are centered on their grand mean, the intercept can still be interpreted as the overall average effect size. Regarding the specific predictors to use, HLMs allow including the reliability of the measures (as an artifact) and the (sample) size of the study as moderators, with the latter being an evaluation of potential publication bias. Moreover, moderators at level-1, level-2 and level-3 can be modeled which is not the case in traditional meta-analysis. Examples of moderators include participant (level-2) characteristics such as severity, duration, and etiology of the problem, duration of the intervention, and demographic data and study (level-3) characteristics such as characteristics of the setting, type of design used, methodological quality score, and the publication date. Moreover, at level-3, it can be interesting to assess the covariance between baseline level and treatment effect: it is like a moderator analysis using the initial baseline level as a moderator (e.g., the intervention might be less effective for studies having a high baseline level). All these predictors are called "time-invariant" (Shaw & Liang, 2012) because they remain the same throughout the measurement occasions (level-1 units).

Note that if REML is used for estimation (e.g., in order to estimate with higher precision the between-study variance; Hox, 2010), it is not possible to test the effect of moderator variables using the chi-square test comparing the deviances of the models with and without the moderator. The importance of the moderators can be assessed by constructing confidence intervals or comparing models via AIC and BIC.

Assumptions, Modeling Flexibility, and Sensitivity Analysis

The main assumptions of HLMs are: (a) independent and normally distributed level-1 residuals; (b) independent and normally distributed level-2 residuals; (c) independence between the residuals at the different levels; (d) the level-1 variance is the same for all level-2 units and (e), by extension, if there is a third level, the level-2 variance is the same for all level-3 units

HLMs are flexible enough to handle situations in which continuous outcomes or a normal distribution cannot be expected. Specifically, the outcome of interest could be measured as a frequency of occurrence and a HLM would still be applicable (Shadish, Kyse, & Rindskopf, 2013). Additionally, the modeling flexibility is expressed in the possibility to model different data patterns at level-1 (presence or absence of general trend which could be linear or not; an immediate and sustained change versus change in slope), different variance components including the covariance between initial baseline level and treatment effect at level 2 (and at level 3), different error structures including homogeneous or heterogeneous autocorrelation within a case across the A and B conditions, and homogeneous or heterogeneous variance within a case across the A and B conditions.

In the SCED context, autocorrelation requires specific attention due to the amount of discussion (e.g., Busk & Marascuilo, 1988; Huitema, 1985; Sharpley & Alavosius, 1988) on whether measurements (level-1 units) are independent or serially related ("autocorrelated"), with recent reviews showing that autocorrelation is common (Shadish & Sullivan, 2011; Solomon, 2014). Despite the fact HLMs assume independent data, it is possible to extend the basic models by modeling autocorrelation in several different ways (Baek & Ferron, 2013). The most common way of dealing with autocorrelation is by including a first-order autoregressive parameter in the model, which specifies that each measurement is partially dependent on the immediately previous measurement, plus some random error (i.e., unexplained variability). In relation to

serial dependence, the uncertainty present has to be highlighted: whether there is or is not autocorrelation, how to model autocorrelation (e.g., linear or non-linear data trends that have not been modeled appropriately can create the appearance of serial dependencies: Shadish, Kyse, et al., 2013), and whether autocorrelation can be expected to be estimated precisely with small samples (Shadish, Rindskopf, Hedges, & Sullivan, 2013). However, when combining effect sizes, researchers are mainly interested in the overall average treatment effect estimates and the between-case variance and between-study variance. The misspecification (or non-specification) of autocorrelation does not bias these estimates. However, the standard errors might be biased (which is undesirable when performing a meta-analysis). To avoid these issues, robust variance estimation (RVE; Hedges et al., 2010; Moeyaert, Ugille, Ferron, Beretvas, & Van den Noortgate, 2016) can be used for estimating standard errors and confidence intervals around the average treatment effect estimates can be reported.

Also in relation to the variety of possible modeling options, in the meta-analysis of SCED studies, it is recommended to perform a sensitivity analysis. The idea is that the researcher tries several plausible models (e.g., with or without trend, with or without autocorrelation, homogeneous versus heterogeneous variance, etc.) according to the visual inspection of the data and report all the results. These results can be used to check whether the parameter estimates and statistical significance is sensitive to the model. If similar results are obtained across a variety if different plausible models, the researcher can be more confident in the results. Model comparisons can be performed not only using the chi-square test on deviances or the AIC (Akaike) and BIC (Schwartz's Bayesian), which may be problematic for small samples (Moeyaert et al., 2014) but also assessing whether the differences in the relevant fixed effect and variance estimates are relevant from a substantive perspective.

The general idea underlying sensitivity analysis is that the results are always relative to the assumptions the data analyst is willing to make. Therefore, it could be stated that all models are wrong, but some models are more suitable than others. The same logic can be applied for the data analysis in each separate SCED study: given the lack of consensus regarding the optimal analytical approach, the consistency across different effect sizes estimators (i.e., different analytical techniques each potentially focusing on a different data aspect, such as overlap, level, trend) boosts the confidence in the conclusions regarding intervention effectiveness (Kratochwill et al., 2010). Nevertheless, not specifying beforehand the type of effect expected and multiple analyses of the same data have been considered a concern both from an ethical perspective and in relation to the validity of the statistical conclusions (Levin, Ferron, & Gafurov, 2017).

Summary of the Advantages for SCED Data

The following advantages are noteworthy. First, a HLM can be a good way of summarizing the evidence when there are within-study replications required for internal validity (two-level analysis) and when performing meta-analysis required for external validity (three-level analysis). Second, it is possible to account for spontaneous recovery (i.e., to model baseline trend), with the model being either linear or non-linear (e.g., Hembry et al., 2015). Third, beyond general time trend, in terms of the intervention effect, the research can model not only change in level, as in the BC-SMD (Shadish et al., 2014), but also change in slope. According to the data features, the researcher can decide which of these aspects (general trend, immediate change in level, change in slope) should be allowed to vary across participants (Baek et al., 2016). Fourth, a single quantification of the magnitude of effect can be obtained, similar to Cohen's *d* or the BC-SMD (Pustejovsky et al., 2014). Fifth, it is possible to model both continuous response variables (e.g., percentages, rates) and counts (e.g., frequency of occurrence) (e.g., Shadish et al., 2013). Sixth, it

is possible to model a specific error structure according to the data features (autocorrelation, heterogeneous variance across phases for SCED). Seventh, the design matrix can be specified in such a way as to apply multilevel models beyond the multiple-baseline design (Moeyaert, Ugille, Ferron, Beretvas, & Van den Noortgate, 2014; Shadish et al., 2013), which is the common situation in which multilevel models are described (Ferron et al., 2009; Moeyaert et al., 2014). Finally, computer code has been offered in several articles for applying HLMs to SCED data (see the list at https://osf.io/sdv4m/). A recently created user-friendly tool is also available to assist the applied SCED analyst at http://52.14.146.253/MultiSCED/ (Declercq, Cools, Beretvas, Moeyaert, & Van den Noortgate, 2018).

An Integration of Visual Analysis and HLM to Synthesizing the Results of SCED Studies

Moeyaert et al. (2014) illustrated the use of a three-level model for quantitatively integrating the results of five multiple-baseline studies examining effects of pivotal response training with children with autism and measuring the percentage of trials with appropriate speech as an outcome variable. Moeyaert et al. (2014) illustrate several modeling options, the simplest of which represents the intervention effect as an abrupt and sustained change in level (i.e., mean difference) without modeling any time trends. For this simplest model, Moeyaert et al. (2014) report an average increase of 31.07% when comparing the intervention phase level to the baseline phase level, a statistically significant difference ($p < .05$). They also report a between-studies standard deviation in the treatment effect equal to 16.49 (not statistically significant) and a between-case standard deviation in the treatment effect equal to 14.97 (statistically significant). These results can be represented visually using the modified Brinley plot (Blampied, 2017). Each point represents the baseline mean (abscissa) and the intervention phase mean (ordinate) for the same case. Each case within the same study is indicated with the same color. The

modified Brinley plot was constructed for the same data analyzed by Moeyaert et al. (2014) using https://manolov.shinyapps.io/Brinley/ and is represented in Figure 1. The solid diagonal line represents no change between baseline and intervention phase mean, whereas the dashed line here represents the average difference as estimated with the three-level model (31.07). The axes of the Brinley plot are defined by the smallest and largest data point observed in the five studies, considering that in some of them summations were used leading to scores above 100%. It can be seen that there is an increase (i.e., improvement) for all cases and that the (same-color) points for the cases belonging to the same studies are sometimes far away, illustrating the finding of statistically significant variation in treatment effectiveness between cases within the studies. The points (cases) belonging to the different studies do not exactly overlap, i.e., there is variation across studies. However, only the points belonging to some studies form clusters, and there is not a clear separation between studies: this seems well-aligned with the three-level finding that the between-studies standard deviation is not statistically significant.

INSERT FIGURE 1 ABOUT HERE

Evidence on the Performance of Three-Level Models for SCED Data

Criteria

Several simulation studies have been carried out on three-level HLMs in order to tests its performance in terms of: (a) whether the fixed effects and random effects are biased or not (i.e., whether the average of many simulations is equal to the simulation parameters) and precise or not (i.e., mean squared error); (b) whether the confidence intervals include the simulation parameter as many times as expected (e.g., 95%); (c) whether the rate of rejection of the null hypothesis that the population parameter is equal to zero is as expected in absence of effect (i.e.,

5% for estimating Type I error rates) and whether this rate of rejection is high enough in presence of effect (i.e., a statistical power of 80%).

Modeling Change in Level in Raw Data

Owens and Ferron (2012) applied three-level models on raw data modeling change in level (i.e., using only the dummy phase variable as a predictor). They found unbiased average treatment effect (i.e., fixed effect) estimates. The confidence intervals for the fixed effects were excessively wide for 10 studies (level-3 units) and improved for 30 studies. Regarding the random effects, the within-case variances are unbiased, but the between-study variances are underestimated and the between-case variances are overestimated.

Modeling Immediate Change and Change in Slope in Raw Data

Other simulation studies focused on three-level models that model both the immediate effect of the intervention (i.e., the dummy phase predictor) and the effect of intervention on time trend (i.e., change in slope modeled via the interaction term between the session number predictor and the dummy phase predictor). When the three-level models are applied on raw data, unbiased average intervention (fixed) effects are found (Moeyaert, Ugille, Ferron, Beretvas, & Van den Noortgate, 2013b, 2013c; Ugille, Moeyaert, Beretvas, Ferron, & Van den Noortgate, 2012). In terms of statistical significance, in order to have reasonable power (≥ .80) for testing the treatment effects, a homogeneous set of at least 30 studies should be included (Moeyaert et al., 2013b). Regarding the random effects, the between-case variances are biased (Moeyaert et al., 2013b). Specifically, Ugille et al. (2012) found that the between-case variance was seriously overestimated when there were only 10 measurement occasions per case.

Modeling Immediate Change and Change in Slope in Standardized Data

When the three-level models focused on immediate intervention effect and the effect of intervention on time trend are applied to standardized data, biased average intervention effects were found (Ugille et al., 2012), especially when there are fewer than 20 measurements (level-1 units). Complementarily, the intervention effects were not biased when there were 20 or more measurements per case, 30 or more studies (level-3 units), and when the between-study variance in treatment effects is small (Moeyaert et al., 2013c). In order to reduce the small-sample bias in the estimates of the fixed effects, Hedges' correction has been shown to function well for uncorrelated data (Ugille, Moeyaert, Beretvas, Ferron, & Van den Noortgate, 2014).

Regarding the random effects, biased estimates of between-study variance and the between-case variance for the immediate treatment effect are found, but for between-case variance, better results are obtained when there are 40 or more measurements (level-1 units) (Moeyaert et al., 2013c).

Modeling Additional Aspects of the Data

One aspect that can affect the internal validity of SCED studies are external events, also known as "history". If an external event is present, but not modeled, there is bias in the estimates of the treatment effects, especially for a small number of studies (10 level-3 units) and few measurement occasions (15 level-1 units); there is also bias in the variance estimates (Moeyaert, Ugille, Ferron, Beretvas, & Van den Noortgate, 2013a).

Another relevant aspect is the assumption of independence between the residuals at levels 2 and 3. Moeyaert, Ugille, et al. (2016) studied the effect of misspecification of the covariance. They found that the treatment effect estimates across cases and across studies are relatively robust against misspecification of the covariance matrix (i.e., ignoring covariance). However,

this is not the case for the estimates of the between-case and between-study variance. Therefore, to improve the estimation of the random components, it is necessary to include in the model the covariance term.

Apart from the assumption of independence of residuals from different levels, it is also usually assumed that the level-1 residuals are independent. For that reason, Petit-Bois et al. (2016) studied the effect of different specifications of the autoregressive error structure for the measurements. They found that in practically all conditions, the treatment effects are unbiased regardless of whether there is misspecification or not, whereas the estimates of the variance components are biased.

Discussion

Recommendations for Conducting Research

Currently, there is a variety of methodological guidelines or rubrics that suggest how a SCED study should be conducted in order to favor both internal and external validity (e.g., Horner et al., 2005; Reichow et al., 2008; Tate et al., 2013; see also Maggin, Briesch, Chafouleas, Ferguson, & Clark, 2014, and Smith, 2012, for reviews). In the current section, we provide further recommendations highlighting favorable conditions for applying HLMs, but more importantly, these recommendations are also expected to entail methodological improvements.

First, in order to decrease within-case variability for reducing the bias in fixed effect estimates when standardizing the raw scores, we recommend applied researchers to wait for a stable baseline and to reduce measurement error by measuring at the same time of day and in the same conditions and with optimal instruments (e.g., observation with high interobserver agreement). Second, it is desirable to collect at least 20 measurements per case in order to ensure

better conditions for standardizing the raw scores. Actually, 20 is the median series length in published SCED research according to the review by Shadish and Sullivan (2011). Third, it is important to ensure treatment fidelity and procedural fidelity in general (Ledford & Gast, 2014) in order to reduce between-case variability. Fourth, it is important to replicate studies (in similar conditions) to (a) increase internal and external validity; (b) have more studies (level-3 units) to meta-analyze; and (c) potentially reduce between-cases and within-cases variability needed for better estimation of the variance components. Fifth, it is also desirable to replicate varying certain conditions in order to allow for moderator analysis (i.e., a better understanding of the factors related to larger versus smaller effects of the intervention: when is the effect of the intervention optimal).

Recommendations for Using HLMs to Meta-Analyzing SCED Research

In terms of obtaining unbiased results when combining standardized data, it is recommended to apply Hedges' small sample bias correction (Ugille et al., 2014). In terms of interpreting the results, researchers should be cautious when fewer than 30 studies are available, given that such situations are related to less precise fixed effect estimates. In that sense, searching for as many studies as possible is also methodologically desirable in order to reduce the probability of publication bias affecting the overall summary. Additionally, researchers should also be cautious when fewer than 20 measurements per case are available. Therefore, it is advised to (a) focus on the fixed effect estimates and to the results for the moderator analysis rather than to the estimates of the variance components; and (b) make explicit comments that the estimates obtained in an analysis may be biased or not sufficiently precise due to the number of level-3 and/or level-1 units available. If the number of units available for synthesis is far below the desired level, the

researcher could opt for an alternative meta-analytical technique, such as combining BC-SMDs (see Maggin et al., 2017) or using Bayesian estimation techniques (i.e., Moeyaert et al., 2017).

Recommendations for the Roles of Applied Researchers and Methodologists When Performing a Multilevel Meta-Analysis of SCED Data

The aim of the article is to make HLMs conceptually accessible to applied researchers, but this does not necessarily entail that applied researchers are expected to be able to apply this analysis by themselves without any assistance (Pustejovsky & Ferron, 2017). The potential need to specify several models and the specific features of the software to use is likely to call for a collaboration between applied researchers and a methodologist.

The applied researcher would discuss: (a) the relevant information to be obtained by using a multilevel model (e.g., what effect to model: a level change or a combination of immediate change and a change in slope); (b); whether nonlinear data patterns (especially in the intervention phase) are expected and meaningful for the type of target behavior and intervention; (c) the importance of including the covariance between the baseline level and the treatment effect; (d) whether it is expected that the across-phases variance within cases is homogeneous or not; (e) whether the data are expected to include autocorrelation (e.g., according to the target behavior and typical time interval between measurement occasions); (f) what the relevant moderators are expected to be for explaining the variation across cases and/or across studies; (g) whether the results of the different models compared in the sensitivity analysis (not necessarily using a statistical test, AIC or BIC) are sufficiently similar; and (h) considering how to interpret the results in case sensitivity analysis leads to results that are deemed “too different” across alternative models.

The methodologist would: (a) help in planning the bibliographic search (keywords, databases, eligibility criteria) and the coding of relevant information; (b) organize the data in a stacked or "person-period" data file necessary for applying a HLM; (c) advice whether the number of level-1, level-2, and level-3 units seems sufficient for using multilevel analysis; (d) discuss whether standardizing the raw data or using effect sizes is a better option for ensuring comparable outcomes; (e) run the models chosen in the appropriate software; (f) help in the interpretation of the numerical output.

Limitations

The current text aimed to present a verbal description of the features of HLMs and, therefore, the formal mathematical and statistical presentations are available elsewhere (i.e., Appendix). Moreover, regarding the meta-analysis of SCED research, we only covered HLM as an option, but we do not claim that it is the only possibility, or that is the only statistically solid option. Regarding such statistically sound options, Shadish et al. (2014) offer an introduction of meta-analysis using the BC-SMD, whereas several meta-analyses using the BC-SMD are described in a Special Issue (Maggin et al., 2017). Additionally, Solmi and Onghena (2014) and Onghena et al. (2018) discuss the possibility to combine probabilities. Moreover, actual meta-analytical practice could be represented by merely obtaining medians or means of indices with unknown sampling distribution (Maggin et al., 2011; Schlosser, Lee, & Wendt, 2008), which does not allow for following all the steps of a typical meta-analysis. Finally, it should be stressed that the quantitative result of a meta-analytic method should be interpreted considering the methodological quality of the evidence (Pustjovsky & Ferron, 2017).

Lines for Future Research

Regarding future research, after the recent focus on the most common multiple-baseline design, it is also relevant to go deeper in the application of multilevel models for reversal designs, alternating treatment designs, and changing criterion designs. A second important quest is to propose and compare alternative ways to standardize the data. Third, regarding the response variable, it would be relevant to study the performance of multilevel models with continuous but non-normal data and also of Poisson models for count data. Fourth, Bayesian estimation has been compared to maximum likelihood estimation for two-level models (Moeyaert, Rindskopf, et al., 2017), but not for three-level models (Chow & Hoijtink, 2017). Moreover, bootstrapping is another alternative that has to be explored further (Van den Noortgate & Onghena, 2005). Fifth, multivariate HLMs need to be studied for SCED data. Sixth, cross-classified data can also be modeled for HLM and more research is needed on this topic.

References

Baek, E. K., & Ferron, J. M. (2013). Multilevel models for multiple-baseline data: Modeling across-participant variation in autocorrelation and residual variance. *Behavior Research Methods, 45*, 65–74.

Baek, E. K., Petit-Bois, M., Van den Noortgate, W., Beretvas, S. N., & Ferron, J. M. (2016). Using visual analysis to evaluate and refine multilevel models of single-case studies. *The Journal of Special Education, 50*, 18–26.

Beeson, P. M., & Robey, R. R. (2006). Evaluating single-subject treatment research: Lessons learned from the aphasia literature. *Neuropsychological Review, 16*, 161–169.

Beretvas, S. N., & Chung, H. (2008). A review of meta-analyses of single-subject experimental designs: Methodological issues and practice. *Evidence-Based Communication Assessment and Intervention, 2*, 129-141.

Blampied, N. M. (2017). Analyzing therapeutic change using modified Brinley plots: History, construction, and interpretation. *Behavior Therapy, 48*(1), 115-127.

Busk, P. L., & Marascuilo, L. A. (1988). Autocorrelation in single-subject research: A counterargument to the myth of no autocorrelation. *Behavioral Assessment, 10*, 229–242.

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

Chow, S.-M., & Hoijtink, H. (2017). Bayesian estimation and modeling: Editorial to the second special issue on Bayesian data analysis. *Psychological Methods, 22*, 609–615.

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.) (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York, NY: Russell Sage Foundation.

Declercq, L., Cools, W., Beretvas, S. N., Moeyaert, M., Ferron, J. M., & Van den Noortgate, W. (under review). *MultiSCED: A tool for (meta-)analyzing single-case experimental data*. Manuscript submitted for publication.

Dedrick, R. F., Ferron, J. M., Hess, M. R., Hogarty, K. Y., Kromrey, J. D., Lang, T. R. … Lee, R. S. (2009). Multilevel modeling: A review of methodological issues and applications. *Review of Educational Research, 79,* 69–102.

Dixon, M. A., & Cunningham, G. B. (2006). Data aggregation in multilevel analysis: A review of conceptual and statistical issues. *Measurement in Physical Education and Exercise Science, 10,* 85–107.

Duval, S., & Tweedie, R. (2000a). Trim and fill: A simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56,* 455–463.

Duval, S. J., & Tweedie, R. L. (2000b). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95*(449), 89–98.

Duval, S. J. (2005). The trim and fill method. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.) *Publication bias in meta-analysis: Prevention, assessment, and adjustments* (pp. 127–144). Chichester, England: Wiley

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ, 315,* 629–634.

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

Ferron, J. M., Farmer, J. L., & Owens, C. M. (2010). Estimating individual treatment effects from multiple-baseline data: A Monte Carlo study for multilevel-modeling approaches. *Behavior Research Methods, 42,* 930–943.

Gage, N. A., & Lewis, T. J. (2014). Hierarchical linear modeling meta-analysis of single-subject design research. *Journal of Special Education, 48*, 3–16.

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

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. *The Journal of Experimental Education, 83*, 514–546.

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children, 71*, 165–179.

Howick, J., Chalmers, I., Glasziou, P., Greenhaigh, T., Heneghan, C., Liberati, A., et al. (2011). *The 2011 Oxford CEBM Evidence Table* (Introductory Document). Oxford Centre for Evidence-Based Medicine. http://www.cebm.net/index.aspx?o=5653

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York, NY: Routledge.

Huitema, B. E. (1985). Autocorrelation in behavior analysis: A myth. *Behavioral Assessment, 7*, 107–118.

Huitema, B. E., & McKean, J. W. (2000). Design specification issues in time-series intervention models. *Educational and Psychological Measurement, 60*, 38–58.

Jamshidi, L., Heyvaert, M., Declercq, L., Fernández Castilla, B., Ferron, J. M., Moeyaert, M., Beretvas, S. N., Onghena, P., & Van den Noortgate, W. (2017, December 28). Methodological quality of meta-analyses of single-case experimental studies. *Research in Developmental Disabilities*. Advance online publication. https://doi.org/10.1016/j.ridd.2017.12.016

Jenson, W. R., Clark, E., Kircher, J. C., & Kristjansson, S. D. (2007). Statistical reform: Evidence-based practice, meta-analyses, and single subject designs. *Psychology in the Schools*, *44*, 483–493.

Kalaian, H. A., & Raudenbush, S. W. (1996). A multivariate mixed linear model for meta-analysis. *Psychological Methods, 1*, 227–235.

Kennedy, C. H. (2005). *Single-case designs for educational research*. Boston, MA: Pearson.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M. & Shadish, W. R. (2010). Single-case designs technical documentation. Retrieved from What Works Clearinghouse website: https://ies.ed.gov/ncee/wwc/Docs/ReferenceResources/wwc_scd.pdf

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

Kreft, I. G. G., de Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research, 30*, 1–21.

Ledford, J. R., & Gast, D. L. (2014). Measuring procedural fidelity in behavioural research. *Neuropsychological Rehabilitation, 24*, 332–348.

Levin, J. R., Ferron, J. M., & Gafurov, B. S. (2017). Additional comparisons of randomization-test procedures for single-case multiple-baseline designs: Alternative effect types. *Journal of School Psychology, 63*, 13–34.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Maggin, D. M., Briesch, A. M., Chafouleas, S. M., Ferguson, T. D., & Clark, C. (2014). A comparison of rubrics for identifying empirically supported practices with single-case research. *Journal of Behavioral Education, 23*, 287–311.

Maggin, D. M., Lane, K. L., & Pustejovsky, J. E. (2017). Introduction to the special issue on single-case systematic reviews and meta-analyses. *Remedial and Special Education, 38*, 323–330.

Maggin, D. M., O'Keefe, B. V., & Johnson, A. H. (2011). A quantitative synthesis of methodology in the meta-analysis of single-subject research for students with disabilities: 1985–2009. *Exceptionality, 19*, 109–135.

Manolov, R. (2017, August 17). Linear trend in single-case visual and quantitative analyses. *Behavior Modification*. Advance online publication. https://doi.org/10.1177/0145445517726301

Moeyaert, M., Ferron, J., Beretvas, S., & Van den Noortgate, W. (2014). From a single-level analysis to a multilevel analysis of since-case experimental designs. *Journal of School Psychology, 52*, 191–211.

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

Moeyaert, M., Ugille, M., Beretvas, S. N., Ferron, J., Bunuan, R., & Van den Noortgate, W. (2017). Methods for dealing with multiple outcomes in meta-analysis: A comparison between averaging effect sizes, robust variance estimation and multilevel meta-analysis. *International Journal of Social Research Methodology, 20*, 559–572.

Moeyaert, M., Ugille, M., Ferron, J., Beretvas, S., & 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*, 547–559.

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

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

Moeyaert, M., Ugille, M., Ferron, J., Beretvas, S. N., & Van den Noortgate, W. (2014). The influence of the design matrix on treatment effect estimates in the quantitative analyses of single-case experimental designs research. *Behavior Modification, 38*, 665–704.

Moeyaert, M., Ugille, M., Ferron, J. M., Beretvas, S. N., & Van den Noortgate, W. (2016). The misspecification of the covariance structures in multilevel models for single-case data: A Monte Carlo simulation study. *The Journal of Experimental Education, 84*, 473–509.

Onghena, P., Michiels, B., Jamshidi, L., Moeyaert, M., & Van den Noortgate, W. (2018). One by one: Accumulating evidence by using meta-analytical procedures for single-case experiments. *Brain Impairment, 19*, 33-58.

Owens, C. M., & Ferron, J. M. (2012). Synthesizing single-case studies: A Monte Carlo examination of a three-level meta-analytic model. *Behavior Research Methods, 44*, 795–805.

Parker, R. I., Cryer, J., & Byrns, G. (2006). Controlling baseline trend in single-case research. *School Psychology Quarterly, 21*, 418-443.

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*, 803–812.

Pustejovsky, J. E., & Ferron, J. M. (2017). Research synthesis and meta-analysis of single-case designs. In J. M. Kauffman, D. P. Hallahan, & P. C. Pullen (Eds.), *Handbook of special education* (2nd ed.) (pp. 168-186). New York, NY: Routledge.

Pustejovsky, J. E., Hedges, L. V., & Shadish, W. R. (2014). Design-comparable effect sizes in multiple baseline designs: A general modeling framework. *Journal of Educational and Behavioral Statistics, 39*, 368–393.

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

Reichow, B., Volkmar, F., & Cicchetti, D. (2008). Development of the evaluative method for evaluating and determining evidence-based practices in autism. *Journal of Autism and Developmental Disorders, 38*, 1311–1319.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86*, 638–641.

Schlosser, R. W. (2005). Meta-analysis of single-subject research: How should it be done? *International Journal of Language and Communication Disorders, 40*, 375–378.

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

Shadish, W. R., Hedges, L. V., & Pustejovsky, J. E. (2014). Analysis and meta-analysis of single-case designs with a standardized mean difference statistic: A primer and applications. *Journal of School Psychology*, *52*, 123–147.

Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods, 43*, 971–80.

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., 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.

Sharpley, C. F., & Alavosius, M. P. (1988). Autocorrelation in behavior data: An alternative perspective. *Behavioral Assessment, 10*, 243–251.

Shaw, B. A., & Liang, J. (2012). Growth models with multilevel regression. In J. T. Newsom, R. N. Jones, S. M. Hofer, *Longitudinal data analysis: A practical guide for researchers in aging, health, and social sciences* (pp. 217-242). Hove, East Sussex, UK: Routledge.

Smith, J. D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods, 17*, 510–550.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London, UK: Sage Publications.

Solmi, F., & Onghena, P. (2014). Combining p-values in replicated single-case experiments with multivariate outcome. *Neuropsychological Rehabilitation, 24*, 607-633.

Solomon, B. G. (2014). Violations of assumptions in school-based single-case data: Implications for the selection and interpretation of effect sizes. *Behavior Modification, 38*, 477–496.

Tate, R. L., Perdices, M., Rosenkoetter, U., Wakim, D., Godbee, K., Togher, L., & McDonald, S. (2013). Revision of a method quality rating scale for single-case experimental designs and n-

of-1 trials: The 15-item Risk of Bias in N-of-1 Trials (RoBiNT) Scale. *Neuropsychological Rehabilitation, 23,* 619–638.

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

Ugille, M., Moeyaert, M., Beretvas, S. N., Ferron, J., & Van den Noortgate, W. (2014). Bias corrections for standardized effect size estimates used with single-subject experimental designs. *The Journal of Experimental Education, 82*, 358–374.

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods, 45*, 576–594.

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

Van den Noortgate, W., & Onghena, P. (2005). Parametric and nonparametric bootstrap methods for meta-analysis. *Behavior Research Methods, Instruments & Computers, 37,* 11–22.

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

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*, 1701–1716.

Table 1

Overview Common Nesting Structures for the Meta-Analysis of Single-Case Experimental Designs Data

	Structure 1	**Structure 2**	**Structure 3**	**Structure 4**	**Structure 5**
Measurements within an AB-comparison	Level-1 units	Level-1 units	Level-1 units	NA (e.g., due to insufficient reporting)	NA, but an effect size is available
Several AB-comparisons for an outcome	NA in a multiple-baseline design	Level-2 units as in an ABAB design	Level-2 units as in an ABAB design	Level-1 units: effect size measure for each AB-comparison	NA in a multiple-baseline design
Several outcomes per participant	NA if one outcome of interest	NA if one outcome of interest	Level-3	NA if one outcome of interest	NA in a multiple-baseline design
Several participants per study	Level-2 units	Level-3 units	Level-4 units	Level-2 units	Level-1 units, each with one effect size
Several studies	Level-3 units	Level-4 units	Level-5 units	Level-3 units	Level-2 units

Note. NA – not available.

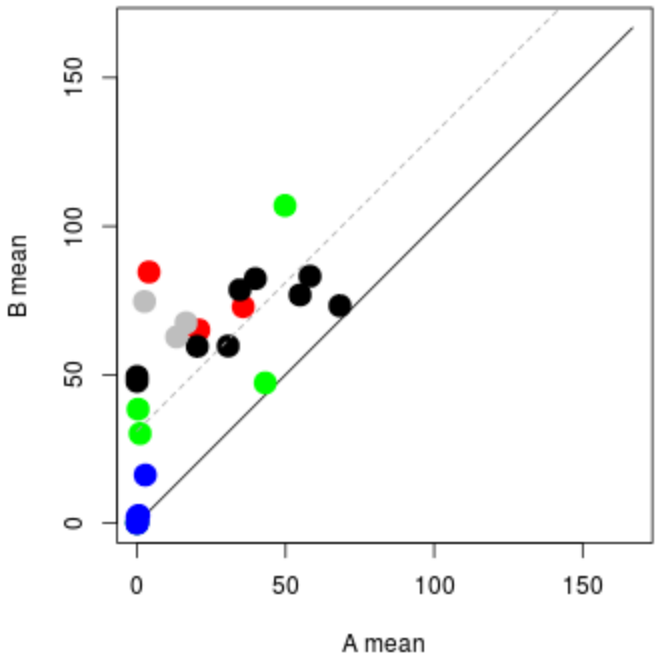

Figure 1. Modified Brinley plot representing the phase A mean (abscissa) and the phase B mean (ordinate) for each participant in the five studies meta-analyzed by Moeyaert, Ferron, Beretvas, and Van den Noortgate (2014). Each color represents a separate study.

Appendix

Recommended textbooks on HLMs and multilevel models, in general, beyond their application to SCED data:

- Goldstein, H. (1995). *Multilevel statistical models*. London, UK: Edward Arnold.
- Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York, NY: Routledge.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models* (2nd ed.). Thousand Oaks, CA: Sage

Articles illustrating how to use multilevel meta-analysis in SCED or using multilevel meta-analysis as part of a broader purpose:

- Baek, E. K., & Ferron, J. M. (2013). Multilevel models for multiple-baseline data: Modeling across-participant variation in autocorrelation and residual variance. *Behavior Research Methods, 45*, 65–74.
- Baek, E., Moeyaert, M., Petit-Bois, M., Beretvas, S. N., Van de Noortgate, W., & Ferron, J. (2014). The use of multilevel analysis for integrating single-case experimental design results within a study and across studies. *Neuropsychological Rehabilitation, 24*, 590–606.
- Dedrick, R. F., Ferron, J. M., Hess, M. R., Hogarty, K. Y., Kromrey, J. D., Lang, T. R. … Lee, R. S. (2009). Multilevel modeling: A review of methodological issues and applications. *Review of Educational Research, 79*, 69–102.

- Gage, N. A., & Lewis, T. J. (2014). Hierarchical linear modeling meta-analysis of single-subject design research. *Journal of Special Education, 48*, 3–16.

- Heyvaert, M., Maes, B., Van den Noortgate, W., Kuppens, S., & Onghena, P. (2012). A multilevel meta-analysis of single-case and small-n research on interventions for reducing challenging behavior in persons with intellectual disabilities. *Research in Developmental Disabilities, 33*, 766–780.

- 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*, 493–510.

- Heyvaert, M., Saenen, L., Maes, B., & Onghena, P. (2015). Comparing the percentage of non-overlapping data approach and the hierarchical linear modeling approach for synthesizing single-case studies in autism research. *Research in Autism Spectrum Disorders, 11*, 112–125.

- Moeyaert, M., Ferron, J., Beretvas, S., & Van Den Noortgate, W. (2014). From a single-level analysis to a multilevel analysis of since-case experimental designs. *Journal of School Psychology*, *52*, 191–211.

- Moeyaert, M., Ugille, M., Ferron, J., Onghena, P., Heyvaert, M., & Beretvas, S. N. (2015). Estimating intervention effects across different types of single-subject experimental designs: Empirical illustration. *School Psychology Quarterly*, *30*, 50–63.

Suggested readings presenting the multilevel models as applied to SCED data, including formulas:

- Moeyaert, M., Ferron, J., Beretvas, S., & Van Den Noortgate, W. (2014). From a single-level analysis to a multilevel analysis of since-case experimental designs. *Journal of School Psychology*, *52*, 191–211.

- Onghena, P., Michiels, B., Jamshidi, L., Moeyaert, M., & Van den Noortgate, W. (2018). One by one: Accumulating evidence by using meta-analytical procedures for single-case experiments. *Brain Impairment*, *19*, 33-58.

- Pustejovsky, J. E., & Ferron, J. M. (2017). Research synthesis and meta-analysis of single-case designs. In J. M. Kauffman, D. P. Hallahan, & P. C. Pullen (Eds.), *Handbook of special education* (2nd ed.) (pp. 168-186). New York, NY: Routledge.

- 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.

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

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

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

Articles reporting simulation studies on the performance of two-level models for analyzing data from SCED studies:

- Ferron, J. M., Bell, B. A., Hess, M. R., Rendina-Gobioff, G., & Hibbard, S. T. (2009). Making treatment effect inferences from multiple-baseline data: The utility of multilevel modeling approaches. *Behavior Research Methods, 41*, 372–384.
- Ferron, J. M., Farmer, J. L., & Owens, C. M. (2010). Estimating individual treatment effects from multiple-baseline data: A Monte Carlo study for multilevel-modeling approaches. *Behavior Research Methods, 42*, 930–943.
- Ferron, J. M., Moeyaert, M., Van den Noortgate, W., & Beretvas, S. N. (2014). Estimating causal effects from multiple-baseline studies: Implications for design and analysis. *Psychological Methods, 19*, 493–510.
- 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. *The Journal of Experimental Education, 83*, 514–546.
- 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*, 175–196.
- Jenson, W. R., Clark, E., Kircher, J. C., & Kristjansson, S. D. (2007). Statistical reform: Evidence-based practice, meta-analyses, and single subject designs. *Psychology in the Schools, 44*, 483–493.
- Moeyaert, M., Rindskopf, D., Onghena, P., & Van den Noortgate, W. (2017, March 30). Multilevel modeling of single-case data: A comparison of maximum likelihood and

Bayesian estimation. *Psychological Methods*. Advance online publication. http://dx.doi.org/10.1037/met0000136

- Moeyaert, M., Ugille, M., Ferron, J., Beretvas, S., & Van Den Noortgate, W. (2013). Modeling external events in the three-level analysis of multiple-baseline across-participants designs: A simulation study. *Behavior Research Methods, 45*, 547-559.
- Mulloy, A.M. (2011). *A Monte Carlo investigation of multilevel modeling in meta-analysis of single-subject research data*. Doctoral dissertation, The University of Texas at Austin, USA. Retrieved from http://hdl.handle.net/2152/ETD-UT-2011-08-3873

Readings on reporting, relevant for meta-analysis and HLMs:

- Ferron, J. M., Hogarty, K. Y., Dedrick, R. F., Hess, M. R., Niles, J. D., & Kromrey, J. D. (2008). Reporting results from multilevel analyses. In A. A. O'Connell, & D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 391–426). Greenwich, CT: Information Age Publishing.
- Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P., ... & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *PLoS Medicine, 6*(7), e1000100.