Behaviormetrika (2021) 48:207–229
https://doi.org/10.1007/s41237-021-00141-z

INVITED PAPER

Assessing generalizability and variability of single-case design effect sizes using two-stage multilevel modeling including moderators

Mariola Moeyaert[1] · **Panpan Yang**[2]

Received: 19 June 2021 / Accepted: 6 July 2021 / Published online: 28 July 2021

Abstract

This study introduces an innovative meta-analytic approach, two-stage multilevel meta-analysis that considers the hierarchical structure of single-case experimental design (SCED) data. This approach is unique as it is suitable to include moderators at the intervention level, participant level, and study level, and is therefore especially recommended for the meta-analyst interested in moving beyond estimating the overall intervention effectiveness. Using this approach, the between-participant variability and between-study variability in intervention effectiveness can be evaluated in addition to obtaining a generalized effect size estimate across studies. This is a timely contribution to the SCED field, as the source(s) of variability in effect size can be identified, and moderators at the corresponding level(s) (participant level and/or study level) can be added to explain the variability. The two-stage multilevel meta-analytic approach, with the inclusion of moderators, can provide evidence-based recommendations about the effectiveness of an intervention taking into account intervention, participant, and study characteristics. First, a conceptual introduction to two-stage multilevel meta-analysis is given to provide a good understanding of its full potentials and modeling options. Second, the usage of this approach will be demonstrated by applying it to a published meta-analytic data set. The goal of this study is to disseminate the two-stage multilevel meta-analysis approach in the hope that SCED meta-analyst will consider this methodology in future meta-analyses.

Keywords Two-stage multilevel meta-analysis · Single-case experimental designs · Moderators · Effect sizes

Communicated by Maomi Ueno.

✉ Mariola Moeyaert
mmoeyaert@albany.edu

Extended author information available on the last page of the article

1 Introduction

Meta-analysis was first introduced in the Social and Behavior Sciences by Gene Glass at the Annual meeting of the American Educational Research Association in 1976 (Glass 1976). Since its introduction, meta-analysis has been widely recognized as a powerful statistical analytic technique to summarize research evidence across studies (Borenstein et al. 2009a, b; Card 2016; Cooper 2017; Hedges and Olkin 1985; Lipsey and Wilson 2001; Sutton et al. 2000). Meta-analysis is one subtype of research synthesis and should not be confused with the other subtypes such as narrative research review, informal vote counting (tallying significance), and formal vote counting (statistical analysis of significance) (Card 2016). Meta-analysis is the statistical analysis of effect sizes (see Card 2016 for an in-depth discussion of the distinction between research synthesis subtypes). The goal of conducting a meta-analysis is to provide a complete overview of (published and unpublished) research evidence, meeting specific inclusion and exclusion criteria, related to a specific topic. In contrast to decision-making at the primary study level, meta-analysis can be used to provide more generalizable, precise, valid, and unbiased conclusions across all identified studies, and can provide explanations for variability in research evidence between studies through inclusion of moderators (Borenstein et al. 2009a, b; Van den Noortgate and Onghena 2008). It is informative to identify under which specific study design conditions an intervention is proven effective. As the number of research reports, publications, conference presentations, dissertations, etc. keeps on increasing exponentially, it is practically impossible for practitioners (e.g., politicians, clinicians, interventionists, teachers, and researchers) to read all available evidence, and as such, meta-analysis is needed and will continue playing a crucial role.

Current study introduces a promising and innovative meta-analytic approach that can be used to synthesize effect sizes obtained from primary studies using a single-case experimental design (SCEDs). SCEDs are unique as these designs repeatedly gather observations for each study participant prior to the start of an intervention (i.e., baseline condition), and during/after the intervention (i.e., intervention condition). In that way, each participant serves as its own control (i.e., no matched comparison group is needed), and individualized data patterns can be observed (Lobo et al. 2017; Moeyaert et al. 2014). An example of typical SCED data is graphically displayed in Fig. 1. The raw data to create the graphical display were retrieved from a published SCED study (Saddler et al. 2017). The software program WebPlotDigitizer was used to retrieve the raw data from the graph displayed in Saddler et al. (2017) and the Shiny tool scdhlm (Pustejovsky et al. 2021) was used to recreate the graph. Saddler and colleagues examined the effects of a summarizing strategy intervention on the quality of written summaries of children with emotional and behavior disorders. The six study participants were repeatedly measured during a baseline condition (i.e., prior to the intervention) and during the intervention condition. This demonstrates the multilayered data structure of SCED studies; repeated observations (i.e., Level 1 = observation or measurement level) are nested within participants (i.e., Level 2 = case or participant level).

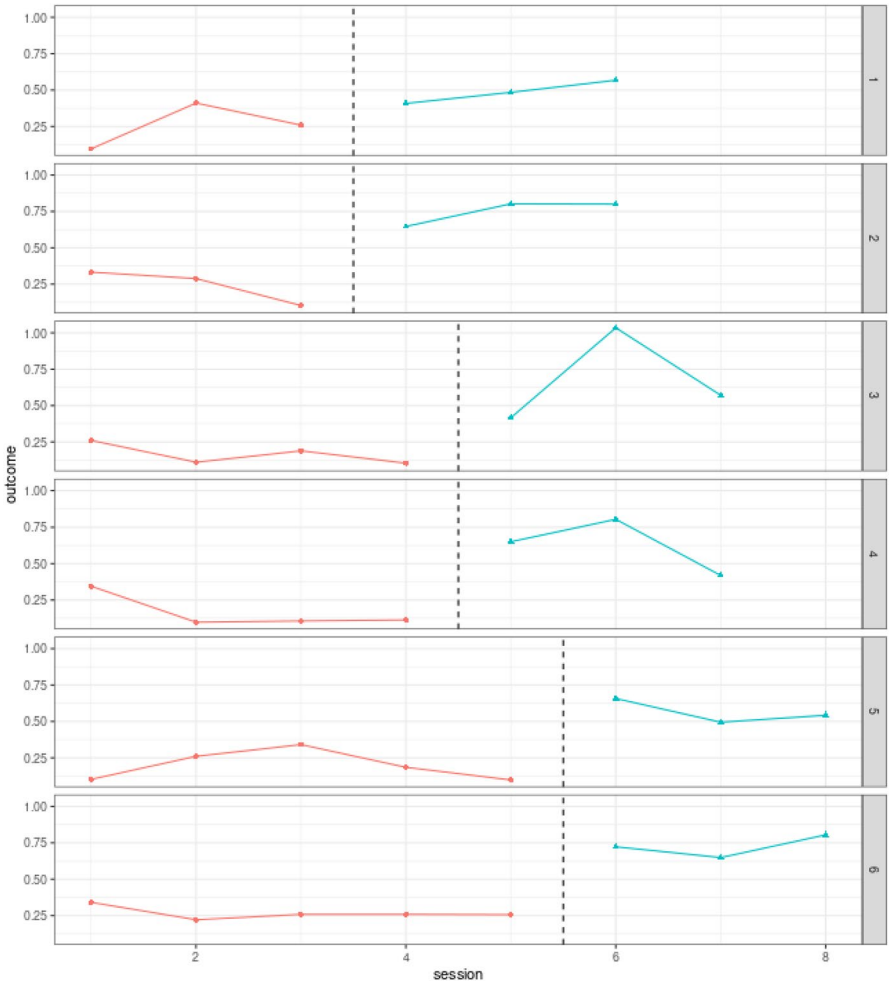

Fig. 1 Graphical display of data from a typical single-case experimental design study (Saddler et al. 2017)

Because of its unique features, the usage of SCEDs to investigate intervention effectiveness is increasing exponentially (see Fig. 2a) in a variety of different fields such as rehabilitation, neurosciences, clinical psychology, and special education (see Fig. 2b). As such, there is a need to quantitatively synthesize the findings across primary SCED studies to identify evidence-based interventions, and make recommendations to the field. Together with the exponential growth, there is a growing demand for methodological sound meta-analytic techniques that can summarize findings from these type of studies. This allows to make inferences and decisions that are based on scientific evidence, which in turn informs practice, theory, and policy.

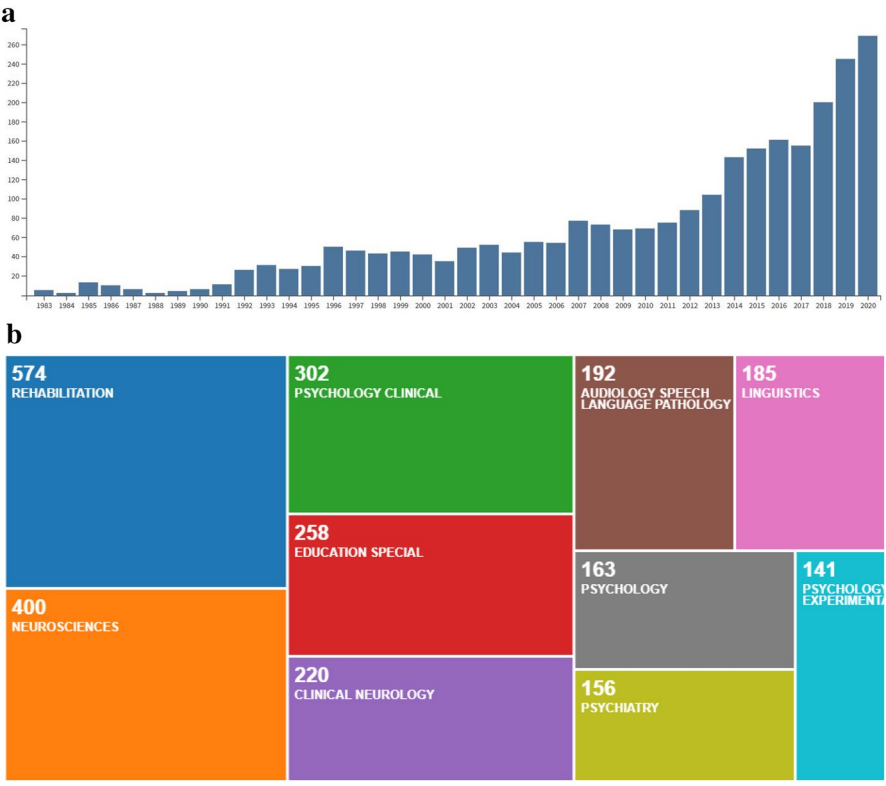

Fig. 2 **a** Overview of the Number of Published SCED Articles over Time (1983–2020) using the Web of Sciences, Keywords: TOPIC = (single-case experiment* OR single-subject experiment). **b** Overview of the number of published SCED articles for the 10 most popular fields, using the Web of Sciences, Keywords: TOPIC = (single-case experiment* OR single-subject experiment)

As the synthesis of SCED effect sizes is relatively new ground, we will first provide a brief introduction into SCED statistics. Then, we will transition into discussing an innovative meta-analytic technique, two-stage multilevel meta-analysis that can be used to summarize SCED statistics across studies. This technique considers the multilayered SCED meta-analytic data structure. Multilevel meta-analysis is the recommended meta-analytic technique as moderators related to the intervention, the participant, and study can be modeled.

1.1 Single-case statistics

In preparation to run a meta-analysis, two summary statistics need to be retrieved or calculated from primary study information: (1) a summary statistic reflecting the size of the effect (preferably expressed on a standardized scale) and (2) a summary statistic reflecting the precision (Borenstein et al. 2009a, b; Card 2016; Cooper 2017; Hedges and Olkin 1985; Lipsey and Wilson 2001; Sutton et al. 2000). These

summary statistics can be directly retrieved if reported in the primary study or can be calculated by plugging in information into algebraic formulas (e.g., means, standard deviation, and standard deviations) programmed into specialized online calculators (e.g., web-based effect sizes calculator by Wilson: https://www.campbellcollaboration.org/research-resources/effect-size-calculator.html) or software programs (e.g., Review Manager 5, RevMan 5). In a next step, the effect sizes (i.e., summary statistics) can be combined using a fixed or random effects meta-analysis in which the precision (e.g., the inverse of the squared sampling variability) is traditionally used a weight (i.e., the lower the precision, the lower the weight assigned to the study, Lipsey and Wilson 2001). The effect sizes can take on a variety of different forms and are dependent on the primary study designs, the measurement scale of the dependent variable, and the available information reported in the primary studies. Guidelines and tutorials (e.g., Cochrane Handbook for Systematic Reviews of Interventions by Higgins et al. 2021; What Works Clearinghouse Procedures Handbook 2020), online calculators (e.g., Meta-Analysis Effect Size Calculator by Wilson), and specialized software programs (e.g., Review Manager 2020 and Borenstein et al. 2013) have been developed to assist in selecting, calculating, and reporting effect sizes and their precision for a variety of different design types including group-comparison studies and observational studies. These resources, however, do not include the option to select the design: single-case experiment. In addition, major organizations such as the Cochrane and the Campbell collaboration provide specific trainings and materials to calculate effect sizes and precision for designs other than single-case experimental designs. Therefore, a brief overview of SCED statistics is provided first as this is needed to transition to meta-analysis of SCED effect sizes.

1.2 Non-overlap statistics

Traditionally, intervention effectiveness is expressed in the form of the amount of overlap between baseline and intervention data, expressed as a percentage. The most popular one is the percentage of non-overlapping data (PND; Parker et al. 2011; Scruggs et al. 1987). The minimum (or maximum) baseline datapoint is extrapolated into the intervention phase, and the number of data points in the intervention condition below (or above) this extrapolated point is counted. The proportion of intervention data points below the extrapolated baseline point reflects the effectiveness of the intervention. Variations of this non-overlap statistic have been developed over the years to improve this non-overlap statistic. For instance, the percentage of data exceeding the median (PEM; Ma 2006; Parker et al. 2011) extrapolates the median of all baseline data points into the intervention condition. Other non-overlap indices, such as the percentage of all non-overlapping data (PAND; Parker et al. 2007, 2011), the percentage of non-overlap of all pairs (NAP; Alresheed et al. 2013; Parker and Vannest 2009; Parker et al. 2011), and TauU (Fingerhut et al. 2021; Parker et al. 2011), have been developed to avoid relying on just one extrapolated data-point to make decisions about intervention effectiveness. Instead, all data points from the baseline phase and the intervention phase are in a specific matter pairwise compared.

An in-depth overview and discussion of non-overlap statistics can be found in Parker et al. (2011). These non-overlap statistics do not reflect the sizes of the effect (only percentages are listed), and were not developed with respect to a sampling distribution that has desirable statistical properties. Previous meta-analyses using non-overlap indices traditionally calculate the unweighted mean or median non-overlap statistic across all studies and report the range and quartiles (Jamshidi et al. 2021). For that reason, non-overlap statistics are not the best choice to be combined in a meta-analysis as appropriate weighting is lacking. In addition, the magnitude of the size of the intervention effect is missing and as such it is challenging to infer clinical significance.

1.3 Regression-based statistics

In contrast to the non-overlap statistics, regression-based statistics are considered to be "true" effect sizes (APA, 7th edition). The regression-based statistics express the magnitude of the intervention effect, and have a well-established sampling distribution with desirable statistical properties. As such, a measure of precision is obtained which is needed to appropriately weight the contribution of individual effect sizes to the overall effect sizes estimate across studies. The appropriate statistical model to quantify the difference in baseline mean and intervention mean can be expressed as follows:

$$\mathrm{Y_i} = \beta_0 + \beta_1\ \mathrm{Phase_t} + \ \mathrm{e_t}, \tag{1}$$

where $\mathrm{e_t}$ is an independent Gaussian error term with mean 0 and variance σ_e^2. Let $\mathrm{Phase_t}$ be an indicator variable for the phase of the experiment, with 0 denoting the baseline and 1 denoting the intervention. Let $\mathrm{t} = 1, 2, \ldots, \mathrm{T}$ be index time, and $\mathrm{Y_t}$ be the outcome variable observed at time t. By running this regression model, β_1 represents the unstandardized mean difference between baseline and intervention outcome level. As such, β_1 reflects the effect size, and the inverse of its standard error reflects the precision. These summary statistics can be obtained for each of the participants and used as input for the meta-analysis. The simple OLS regression model (Eq. 1) can be extended to a piecewise regression model. This allows to model a trend line in the baseline condition, which can be extrapolated into the intervention. A different time trend can be modeled in the intervention phase (interaction between time and phase). The intervention effect can be conceptualized as the expected difference in outcomes using the extrapolated baseline trend and the actual estimated intervention trend at a chosen point in the intervention phase: $\beta_{1,t} = \mathrm{E(Y|T = t, Phase = 1) - E(Y|T = t, Phase = 0)}$. The estimated effect size at intervention session 1, 2, and 3 is graphically displayed in Fig. 3, using the data from Participant 6 (see Fig. 1) from Saddler et al. (2017). Similarly, the regression coefficient estimate at a particular point into the intervention phase and its precision can be used as input in the meta-analysis.

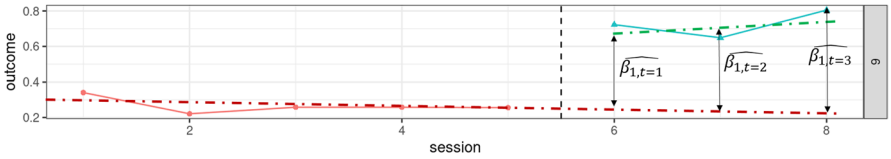

Fig. 3 Graphical display intervention effect size using the piecewise regression model. Data are displayed for one participant from Saddler et al. (2017)

1.4 Multilevel meta-analysis of single-case experimental data

When meta-analyzing SCED studies including multiple participants, three hierarchical levels can be distinguished: repeated measurements are nested within participants, which are nested within studies (Van den Noortgate and Onghena, 2003a, b, 2007, 2008). Traditional meta-analytic techniques in which a summary statistic and precision at the study level are used as input of the meta-analysis are not appropriate as dependency between effect sizes within a study is ignored. Previous methodological research by Van den Noortgate et al. (2005) indicated that ignoring a level (and as such ignoring a source of dependency) results in too small standard errors (i.e., over estimating the precision) of the estimated effect size across studies and as such inflated Type I errors are obtained (i.e., false positive: falsely concluding that an intervention is statistically significant). Therefore, Van den Noortgate and Onghena (2008) introduced the three-level meta-analytic model. To prepare for the SCED meta-analysis, an effect size (together with its standard error/precision) for each of the participants within a primary SCED study needs to be calculated. A standardized mean difference as effect size for each of the participants (which is different from Glass' Δ, Cohen's d, and Hedges' g from group design studies) can be obtained from a regression coefficient (β_1 in Eq. 1). This is possible as raw data for each of the participants can be retrieved from the time-series graphs traditionally displayed in SCED studies (see Fig. 1 for an example). Specialized data retrieval software programs (e.g., WebPlotDigitizer, Ungraph, DataThief, and XYit) can be used for this purpose (see Moeyaert et al. 2016 for details about these programs and the data retrieval process). The raw SCED data can be used as input to run the regression model per participant (Eq. 1) and obtain an estimate of the regression coefficient reflecting the difference in means between baseline and intervention outcome level. However, the obtained regression coefficient is not on a standardized scale, which is recommended, as the scale of the outcome is unlikely to be the same across participants and studies being aggregated. Van den Noortgate and Onghena (2008) introduced a standardization method which was later empirically validated by Ugille et al. (2012) and Moeyaert et al. (2013a). The standardized effect size is obtained by dividing the estimated regression coefficient, $\widehat{\beta}_1$ by the estimated within-participant residual standard error, $\widehat{\sigma}_e$: $b_1 = \frac{\widehat{\beta}_1}{\widehat{\sigma}_e}$. Subsequently, the estimated standardized mean difference and standard error for each of the participants can be used as input to run a three-level meta-analysis. Because this approach involves (1) a pre-processing stage and (2) a meta-analytic stage, it can be best understood as a two-stage

multilevel meta-analysis (2-Stage MLM) or SCED data. For a detailed introduction to 2-Stage MLM, we refer the reader to Declercq et al. (2020). The possibility of including moderators to explain heterogeneity in effect sizes was not discussed by Declercq et al. (2020), and will be introduced in this study. Before transitioning to 2-Stage MLM with the inclusion of moderators, we will provide a discussion of moderators typically encountered in context of SCED meta-analyses in the social and behavior sciences.

1.5 Single-case meta-analyses and moderators

Previous systematic reviews of SCED meta-analyses (Moeyaert et al. 2021a, b; Jamshidi et al. 2021) indicate that there is an interested in explaining heterogeneity in SCED effect sizes estimates between participants and between studies through exploring moderators. SCED meta-analyses commonly report moderators related to the intervention (e.g., dosage if the intervention), participants (e.g., disability status), and/or primary SCED studies (e.g., study quality). A complete overview and description of moderator characteristics can be found in Moeyaert et al. (2021a, b). The systematic review (Moeyaert et al. 2021a, b) report that the most commonly discussed intervention-level moderator is intervention program (e.g., video modeling program versus visual cueing program), the most frequently used participant-level moderator is participant's age, and the most commonly encountered study-level moderator is study design (e.g., multiple baseline design, reversal design, and alternating treatment design). The most frequently used measurement scale of moderators at all three levels is nominal.

As moderator characteristics are commonly reported in primary studies, there is an opportunity to run moderator analyses at the meta-analytic level. The systematic review of Jamshidi et al. (2021) found that 73% of the 178 SCED meta-analyses they reviewed (published between 1985 and 2015) did a moderator analysis. The majority of these meta-analyses simply reported the average effect size per level of the moderator. For instance, the average PND is calculated for male and female participants separately. Only 10% of the SCED meta-analyses applied multilevel analysis to synthesize raw SCED data with the inclusion of moderators. None of these studies used two-stage MLM which is the approach recommended in current study. A subsequent systematic search (Moeyaert et al. 2021a, b) was conducted by replicating the process by Jamshidi et al. (2021) to investigate SCED meta-analysis published after 2015 (until 2020). The search found that 41 meta-analyses of SCEDs discussed and analyzed moderators, while only five SCED meta-analysis used multilevel modeling to summarize the SCEDs including moderators.

Taken together, the importance of analyzing moderators in meta-analyses of SCEDs has been largely recognized. However, most of the existing SCED meta-analyses examined the moderators by aggregating moderators, and reporting average effect sizes per moderator level. Consequently, heterogeneity in effect sizes between participants and between studies remains unexplored. To address this issue, a meta-analytic approach is needed that accounts for the hierarchical SCED meta-analytic data structure, with the option to include moderators. As such, moderators

at their appropriate level (observation level, participant level, and study level) can be modeled accordingly. The three-level modeling approach, as introduced by Van den Noortgate and Onghena (2008) and empirically validated by Moeyaert et al. (2013a, b), is recommended. For a detailed systematic introduction to the basics of three-level multilevel modeling of SCED studies, we refer the reader to Moeyaert et al. (2014). For an extension of the basic model in which one-stage versus two-stage multilevel meta-analysis of SCED studies is discussed, we refer to Declercq et al. (2020). In this study, we will extend the two-stage multilevel meta-analysis to accommodate moderators at the observation (i.e., intervention), participant, and study levels.

2 Methodology

2.1 Two-stage multilevel meta-analysis: unconditional model

The unconditional model is also known as the baseline model, intercepts only model or the model not including any moderators (Raudenbush and Bryk 2002). This model estimates the total amount of variability in effect size estimates, and the amount of variability at the participant and study level. This informs whether there is a need to explore moderators, and at which level of the model the moderators are needed. The moderators might be able to explain variability in obtained effect sizes at the participant and/or study level. The pre-processing step (stage 1 of the 2-stage MLM) provides an estimate of the participant-specific standardized regression coefficient reflecting the effect sizes, b_{1jk}, and the within-participant residual standard deviation, σ_{r1jk}. The following simple ordinary least square regression model can be used for this purpose:

Pre-processing model:

$$y_{ijk} = b_{0jk} + b_{1jk}Phase_{ijk} + r_{1jk} \text{ with } r_{ijk} \sim N(0, \sigma_r^2); \tag{2}$$

y_{ijk} indicates the outcome score at measurement occasion i for participant j who is nested within study k. $Phase_{ijk}$ is a dummy variable indicating whether y_{ijk} is an intervention observation ($Phase_{ijk} = 1$) or intervention observation ($Phase_{ijk} = 1$). Therefore, b_{1jk} indicates the intervention effect. Next, b_{1jk} is a function of the true participant-specific effect size β_{1jk} and the residual standard deviation is assumed to be known (obtained from the pre-processing step in Eq. 2):

Level 1—observation level:

$$b_{1jk} = \beta_{1jk} + r_{1jk}. \tag{3}$$

Next, the participant-specific population effect sizes, β_{1jk}'s, are assumed to vary between participants as it is unlikely that the intervention effect is identical across participants. The participant-specific effect sizes are a function of the study-specific effect sizes (θ_{10k}'s) and a participant-specific deviation (u_{1jk}) from the study-specific effect size. With other words, the effect size for participant j within study k depends on the overall effect size across all participants nested within study k (θ_{10k}), and the

deviation of participant j from the overall effect size (u_{1jk}). The deviations are assumed to be normally distributed with a variance of $\sigma^2_{u_{1jk}}$ (i.e., the between-participant variance in effect sizes).

Level 2—participant level:

$$\beta_{1jk} = \theta_{10k} + u_{1jk} \text{ with } u_{1jk} \sim N\left(0, \sigma^2_{u_{1jk}}\right). \tag{4}$$

Similarly, the study-specific effect sizes are likely to vary between studies and therefore a third level is needed. γ_{100} is the overall effect sizes across all studies. Study-specific effect sizes, θ_{10k}'s, are a function of this overall effect sizes and a study-specific deviation (v_{10k}). The deviations are assumed to be normally distributed with a variance of $\sigma^2_{v_{10k}}$ (i.e., the between-study variance in effect sizes).

Level 3—study level:

$$\theta_{10k} = \gamma_{100} + v_{10k} \text{ with } v_{10k} \sim N\left(0, \sigma^2_{v_{10k}}\right). \tag{5}$$

The research synthesis is interested in the estimate of (1) γ_{100}, reflecting the effectiveness of the intervention across all participants and all studies, and (2) $\sigma^2_{v_{10k}}$ and $\sigma^2_{u_{1jk}}$ indicating the amount of variability in intervention effectiveness between studies and participants, respectively.

2.2 Two-stage multilevel meta-analysis: conditional model

A conditional two-stage multilevel meta-analytic model can be built in an effort to explain heterogeneity in intervention effectiveness. The multilevel meta-analytic approach is recommended for the synthesis of SCED studies as this approach considers the uniqueness of SCED studies: observations are nested within participants and participants are nested within studies. This allows to estimate heterogeneity in effect sizes between participants, and to model participant moderators instead of using aggregated moderators at the study level (e.g., Zelinsky and Shadish 2018). It is important to differentiate between the different levels, as an intervention can be large and statistically significant at the study level, but highly variable at the participant level. This indicates that the intervention is not effective for all participants, and making inferences and recommendations about intervention effectiveness while ignoring individual differences is problematic. There is a need to identify for whom is this intervention working, and under which conditions. If the unconditional model provides evidence for heterogeneity at the participant and/or study level, promising moderators (based on previous research/practice) can be considered at the appropriate level. The level-2 equation can be extended to model a participant moderator in an effort to explain variability in intervention effectiveness between participants. For instance, a researcher might add the moderator gender to the participant level as previous research evidence suggests that the intervention is more successful for female compared to male participants:

Level 2—participant level:

$$\beta_{1jk} = \theta_{10k} + \theta_{11k} Female_{11k} + u_{1jk} \text{ with } u_{1jk} \sim N\left(0, \sigma^2_{u_{1jk}}\right). \quad (6)$$

Gender can be coded as a dichotomous variable, Female in Eq. (5), equaling 0 for male and 1 for female study participants. θ_{10k} indicates the intervention effect for male participants in study k, θ_{11k} reflects the difference in intervention effectiveness between male and female study participants in study k, and θ_{10k}+θ_{11k} is the intervention effect for female participants in study k. The estimated between-participant variance, $\sigma^2_{u_{1jk}}$, obtained by the unconditional model and the conditional model can be compared to evaluate whether the moderator gender decreased the amount of heterogeneity at the participant level. Similarly, moderators at the third level can be considered to explain heterogeneity in effect sizes between studies. For instance, a researcher might be interested in adding study quality as a study-level moderator (based on previous evidence in the field, it is assumed that lower quality report higher effect sizes):

Level 3—study level:

$$\theta_{10k} = \gamma_{100} + \gamma_{101} Quality_{101} + v_{10k} \text{ with } v_{10k} \sim N\left(0, \sigma^2_{v_{10k}}\right). \quad (7)$$

Following the recommendations by the What Works Clearinghouse standards (WWC 2020) for SCEDs, study quality can be coded as (1) not meeting the quality standards, (2) meeting the standards with reservations, and (3) fully meeting the standards. According to the WWC standards, only SCED studies meeting the standards with reservations and fully meeting the standards should be considered for inclusion in the meta-analysis. Instead of excluding the studies not meeting the standards, a dummy-coded study-level moderator can be added with 0 indicating not meeting the standards and 1 reflecting studies meeting the standards (with or without reservation). γ_{100} indicates the intervention effect across all studies for low-quality studies, γ_{101} reflects the difference in intervention effectiveness between low- and high-quality studies, and γ_{100}+γ_{101} is the intervention effect for high-quality studies. The estimated between-study variance, $\sigma^2_{v_{1jk}}$, between the unconditional model and the conditional model can be compared to evaluate whether the moderator, study quality, decreases the amount of heterogeneity at the study level. The combined three-level multilevel meta-analytic model can be obtained by inserting equations [Eqs. (6) and (7)] in Eq. (3)

$$b_{1jk} = \gamma_{100} + \gamma_{101} Quality_{101} + (\gamma_{110} + \gamma_{111} Quality_{111} + v_{11k}) Gender_{11k} + v_{10k} + u_{1jk} + r_{1jk}. \quad (8)$$

Besides the two main effects ($\gamma_{100}, \gamma_{101}$), cross-level interaction effects of the moderators can be looked at ($\gamma_{110}, \gamma_{111}$). Note that we only discussed coding and modeling of dichotomous moderators as a previous systematic review of moderators for SCED meta-analysis indicates that these are the most commonly used measurement scale moderators (Moeyaert et al. 2021a, b). However, if a nominal moderator with more than two categories is of interest, then this moderator can be recoded into a number of dummy-coded moderators (=total number of categories -1). If a continuous moderator is of interest, then it is recommended to center participant moderators around the study

mean, and to center study-level moderators around the grand mean. For more information about coding moderators, see Raudenbush and Bryk (2002).

For simplicity and didactic purposes, only one intervention effect size of interest is combined across participants and across studies and one moderator at the higher levels is modeled. The level-1, level-2, and level-3 equations can easily be extended by including additional effect sizes. For instance, using a piecewise regression model in the pre-processing stage results in multiple effect sizes (see graphical display in Fig. 3) that can be combined across participants and studies (see Ugille et al. 2012). In addition, more than one moderator at the higher levels can be included. In the current study, we focus on combining one intervention effect size (regression-based standardized mean difference), and we consider multiple moderators at level-2 and one moderator at level-3. The model can easily be extended by modeling P number of participant-level moderators at level-2 and Q number of study level moderators at level-3. Equations (8) and (9) reflect the general equations that can be used to model P number of Z and B refer to the level-2 and level-3 moderators, respectively.

Level 2—participant level:

$$\beta_{1jk} = \theta_{10k} + \sum_{p=1}^{P} \theta_{1pk} Z_{1pk} + u_{1jk} \text{ with } u_{1jk} \sim N\left(0, \sigma^2_{u_{1jk}}\right). \tag{9}$$

Level 3—study level:

$$\theta_{10k} = \gamma_{100} + \sum_{q=1}^{Q} \gamma_{10q} B_{10q} + v_{10k} \text{ with } v_{10k} \sim N\left(0, \sigma^2_{v_{10k}}\right) \tag{10}$$

$$\begin{aligned}
\theta_{11k} &= \gamma_{110} + \sum_{q=1}^{Q} \gamma_{11q} B_{11q} + v_{11k} \text{ with } v_{11k} \sim N\left(0, \sigma^2_{v_{10k}}\right) \\
\theta_{12k} &= \gamma_{120} + \sum_{q=1}^{Q} \gamma_{12q} B_{12q} + v_{12k} \text{ with } v_{12k} \sim N\left(0, \sigma^2_{v_{10k}}\right) \\
&\cdots \\
\theta_{1pq} &= \gamma_{1p0} + \sum_{q=1}^{Q} \gamma_{1pq} B_{1pq} + v_{1pk} \text{ with } v_{1pk} \sim N\left(0, \sigma^2_{v_{10k}}\right).
\end{aligned}$$

The combined model as a combination of the level-1, level-2, and level-3 equations can easily become very complex. In this study, we provide a demonstration of the usage of the two-stage multilevel meta-analytic approach by including one moderator at each of the higher levels. A published meta-analytic data set will be used for this purpose. The goal of this paper is to provide a conceptual introduction to this meta-analytic approach, so that meta-analysts fully understand its potentials.

3 Demonstration and application: usage of two-stage multilevel modeling

The three steps involved in two-stage multilevel meta-analysis of SCED studies are demonstrated using a published meta-analytic data set (Moeyaert et al. 2019). These three steps involve (1) pre-processing, (2) unconditional model, and (3) conditional model. An overview of the obtained parameter estimates together with interpretations in context of the study is provided.

3.1 Introduction of empirical example

Moeyaert et al. (2019) synthesized SCED studies to examine the effectiveness of peer-tutoring interventions on both academic and social-behavior performance for at-risk students and students with disabilities. The study authors used a three-level hierarchical linear model to evaluate the effectiveness of peer-tutoring, and to explain heterogeneity in effect sizes at the participant and study level by including moderators. In their study, the authors combined raw data from primary SCEDs instead of combining effect sizes and as such ran a one-stage multilevel meta-analysis (i.e., the pre-processing step is not included). Declercq et al. (2020) recommend two-stage multilevel meta-analysis to reduce model complexity and avoid convergence issues is multiple moderators are considered. This was not considered by Moeyaert et al. (2019). The participant-level moderators include age and gender, and the study-level moderator is study quality. Moeyaert et al. (2019) found that peer-tutoring interventions have a statistical significant effect on academic (γ= 4.18, SE=1.74, p=0.02) and social-behavior performance (γ = 1.84, SE=0.47, p=0.001) for at-risk students and students with disabilities; and the authors also uncovered that participant-level and study-level moderators can reduce some of the between-participant and between-study variance in the effectiveness of peer-tutoring interventions, although the effects of moderators were not statistically significant (all $ps > 0.05$). The authors acknowledge that lack of statistical significance can be due to lack of statistical power. By combining effect sizes instead of raw data, the meta-analytic model is simplified and as such has more power to identify true moderator effects. We will demonstrate the two-stage multilevel modeling approach using solely the academic outcome scores. Some of the primary studies did not include information related to the moderators' age or gender and as such needed to be excluded from the analysis. The study quality was rated for each of the primary studies by Moeyaert et al. (2019), so all information was available for that moderator. This results in 26 primary studies, with a total of 222 participants, available for the empirical demonstration.

3.2 Pre-processing

As explained in Methods section, a simple OLS regression model is run for each of the 222 study participants separately. The regression coefficients and

the precision are standardized and saved in a separate data set, and are used as input of the multilevel meta-analysis. Figure 4 displays a visualization of the structure of the obtained data set; each row represents the participant-specific effect size, and precision, and a study ID and Case ID (i.e., participant ID) are also assigned. In a next step, the moderators' age, gender, and study quality are merged to this data set. The full data set, including the moderators, can be requested by contacting the first author.

To obtain a better understanding of the magnitude and distribution of the effect sizes, a boxplot is created, which is displayed in Fig. 5. The unweighted mean and median effect size across all studies is 1.46 and 1.20, respectively. The skewness and Kurtosis statistics are 0.67 and 0.90, respectively, which indicates that the distribution of effect sizes does not deviate significantly from normality. The range is 14.52 (min = − 4.53, max = 9.99), and the *SD* is 2.62.

In addition, the distribution of effect size estimates per study is visualized in Fig. 6. This provides preliminary evidence that heterogeneity in effect sizes between studies is to be anticipated as the mean effect size per study varies tremendously. In addition, there is a lot of variability in effect size estimates observed within the studies. This can be deduced by analyzing each of the boxplots displayed in Fig. 6 separately.

3.3 Unconditional model

First, the unconditional multilevel meta-analytic model is run (i.e., baseline or intercepts only model) to investigate (1) the effectiveness of peer-tutoring interventions to increase academic outcomes (i.e., estimate of γ_{100}) and (2) variability in effect size estimates between participants and/or studies (i.e., estimate of $\sigma^2_{u_{1jk}}$ and $\sigma^2_{v_{10k}}$, respectively). The following unconditional model is ran in SAS 9.4 (SAS Institute Inc. 2014) using the PROC MIXED statement: $b_{1jk} = \gamma_{100} + v_{10k} + u_{1jk}$ with $u_{1jk} \sim N\left(0, \sigma^2_{u_{1jk}}\right)$ and $v_{10k} \sim N\left(0, \sigma^2_{v_{10k}}\right)$. Based on previous methodologic work in context of multilevel meta-analysis of SCEDs, the Restricted Maximum-Likelihood estimation procedure is specified, and the degrees of freedom are estimated using the Kenward–Roger approach (Ferron et al. 2010). The estimated standardized intervention effect across all studies equals 1.67 [$\widehat{\gamma}_{100} = 1.67$, $SE = 0.49$, $t(24.5) = 3.41$, $p = 0.0022$]. This indicates that, in general, peer-tutoring increases the academic performance by 1.67 standardized units. However, the effectiveness of the peer-tutoring intervention varies between studies [$\widehat{\sigma}^2_{v_{10k}} = 5.58$, $SE = 1.75$, $Z = 3.18$, $p = 0.0007$] and between participants within studies [$\widehat{\sigma}^2_{u_{1jk}} = 1.72$, $SE = 0.28$, $Z = 6.13$, $p < 0.00001$]. This indicates that some participants might benefit from the intervention, whereas others' academic performance does not increase, or even decreases (which is problematic). Before recommending the peer-tutoring intervention to the broader field, it is important to have a good understanding of who is benefitting from the intervention. This will be explored in the next section.

Study	Case	Effect	Precision
1	1	1.895	0.75305
1	2	2.614	0.72742
1	3	0.256	1.34505
1	4	0.064	1.34505
1	5	-1.816	1.61172
1	6	-0.964	1.56081
1	7	1.555	0.75305
1	8	-1.079	0.72742
1	9	0.638	1.34505
1	10	0.063	1.46552
1	11	-0.332	1.61172
1	12	-0.146	1.56081
1	13	0.81	0.75305
1	14	2.006	0.72742
1	15	-0.916	1.34505
1	16	-0.385	1.46552
1	17	-0.878	1.61172
1	18	-0.998	1.56081
1	19	0.369	0.75305
1	20	-0.003	0.72742
1	21	-1.143	1.34505
1	22	-0.003	1.46552
1	23	0.079	1.61172
1	24	-0.935	1.56081
1	25	0.028	0.75305
1	26	-0.629	0.72742
1	27	-0.901	1.34505
1	28	0.259	1.46552
1	29	2.107	1.61172
1	30	0.391	1.56081
2	1	2.919	0.71691
2	2	0.423	1.07417
2	3	1.631	1.07417
2	4	1.227	1.07417
2	5	0.304	3.04028
2	6	2.383	0.68182

Fig. 4 Results pre-processing step of the two-stage multilevel meta-analysis

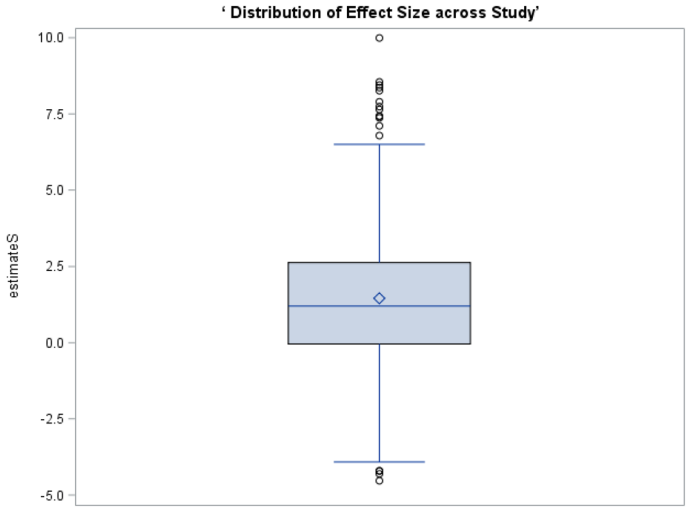

Fig. 5 Distribution of effect size estimates

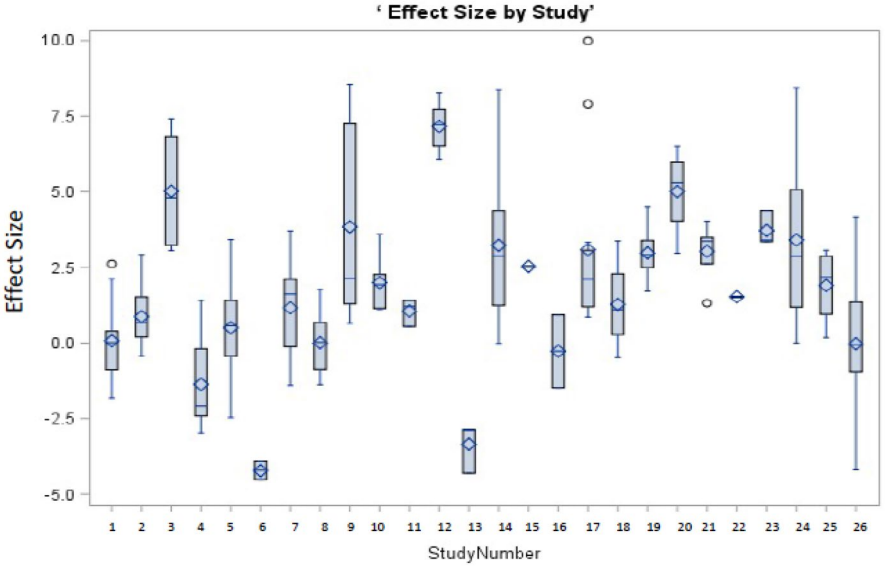

Fig. 6 Distribution of effect size estimates per study

3.4 Conditional model

Similar to the unconditional model, the conditional models are run in SAS 9.4 (SAS Institute Inc., 2014) using the PROC MIXED statement. The restricted maximum likelihood is specified, and the Kenward–Roger method for estimating the degrees of freedom is used.

3.4.1 Participant moderator

Based on the previous research (Moeyaert et al. 2019), it can be assumed that peer-tutoring interventions are likely to be more effective for older children. The average and median age across all 222 participants is 9 and 8, respectively, and the age ranges from 5 to 20. A graphical display of the distribution of age is provided in Fig. 7a. Because some of the age groups have a limited amount of participants and some ages are not included, we first dichotomized the moderator age. Participants younger than 9 are categorized as "young" (age = 0) and participants

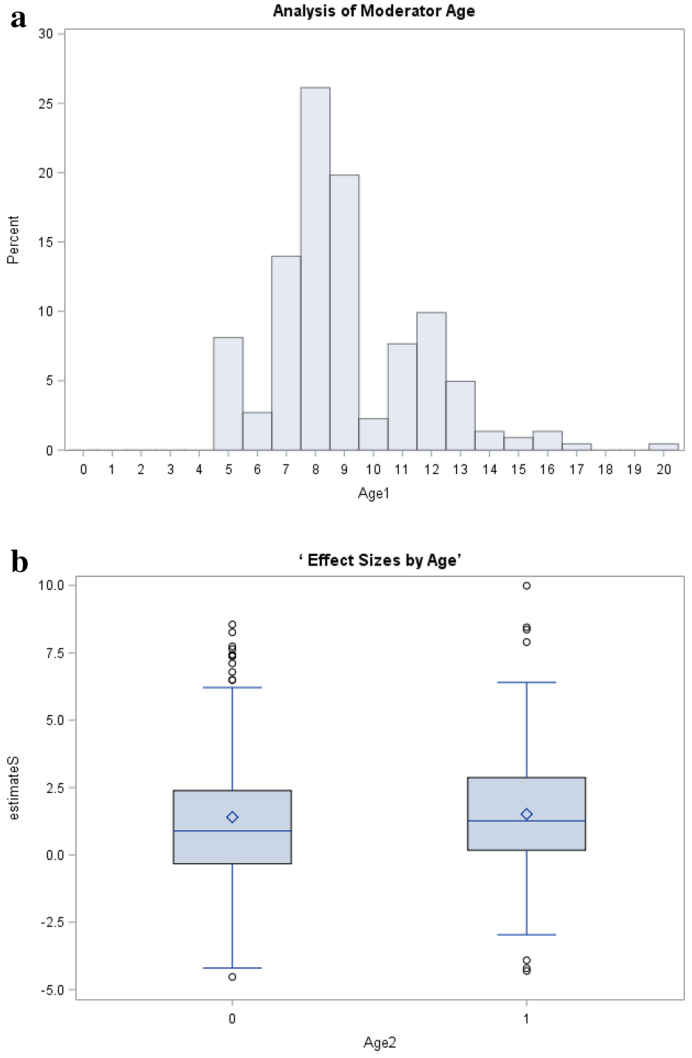

Fig. 7 **a** Distribution of age across the 222 participants. **b** Distribution of effect sizes for younger versus older children

ages 9 or older are categorized as "old" (age = 1). The "young" age group has 113 participants and the "old" age group has 109 participants. Figure 7b provides a graphical display of the distribution of the effects sizes per age group.

For the older age group, the average and median intervention effect equals 1.51 and 1.26, respectively, and varies from − 4.30 to 9.99. For the younger age group, the average and median intervention effect equals 1.40 and 0.88, respectively, and varies from − 4.53 to 8.55. This provides preliminary evidence in support of the hypothesis that peer-tutoring interventions are more effective for older children. However, this preliminary explorative analysis at the primary study level can be misleading as variability within and between studies is not taken into consideration. Unfortunately, this is what has been traditionally done in the previous SCED meta-analyses (Jamshidi et al. 2021). To investigate whether the peer-tutoring intervention has a differential impact on academic outcomes for older versus younger students at the meta-analytic level, age is added as a level-2 moderator, and the following combined meta-analytic model is run: $b_{1jk} = \gamma_{100} + \gamma_{200} Age_{11k} + v_{10k} + u_{1jk} + r_{1jk}$. γ_{100} reflects the intervention effect for young children, and γ_{200} indicates the difference in intervention effectiveness between younger and older children. In addition, it can be evaluated whether the estimated between-participant variability in intervention effectiveness, $\widehat{\sigma}^2_{u_{1jk}}$, decreases with the addition of the moderator. The results indicate that the intervention has a significant impact on academic outcomes for young children [$\widehat{\gamma}_{100} = 2.06$, $SE = 0.61$, $t(40.6) = 3.37$, $p = 0.0016$], and there is no statistically significant difference in intervention effectiveness between younger versus older participants [$\widehat{\gamma}_{200} = -0.67$, $SE = 0.63$, $t(143) = -1.05$, $p = 0.29$]. The effectiveness of the intervention for older children is estimated to be 2.06–0.67 = 1.39. By adding the moderator age as a dichotomous variable, the estimated between-participant variability ($\widehat{\sigma}^2_{u_{1jk}}$) is not reduced and remains around 1.72.

Dichotomizing a continuous variable is not recommended as this changes the measurement scale of the variable, and omits information. Therefore, we re-ran the meta-analysis by including age as a continuous variable. Because age is continuous variable (expressed in years), it is centered around the study average age (per recommendation of Raudenbush and Bryk 2002). Therefore, γ_{100} reflects the intervention effect for students at the average study age, and γ_{200} indicates the change in effectiveness of the intervention between students being one year apart in age. In addition, it can be evaluated whether the estimated between-participant variability in intervention effectiveness, $\widehat{\sigma}^2_{u_{1jk}}$, decreases with the addition of the moderator age as a continuous variable. The results indicate that the intervention has a significant impact on academic outcomes for students at the average study age [$\widehat{\gamma}_{100} = 1.52$, $SE = 0.50$, $t(26.7) = 3.04$, $p = 0.0053$], and that the intervention is more effective for older participants, although this is not statistically significant [$\widehat{\gamma}_{200} = 0.14$, $SE = 0.08$, $t(220) = 1.74$, $p = 0.08$]. The effectiveness of the intervention for students 1 year older compared to the study average is estimated to be 1.52 + 0.14 = 1.66. By adding the moderator age as a continuous variable, the estimated between-participant variability ($\widehat{\sigma}^2_{u_{1jk}}$) becomes almost 20 times smaller compared to the baseline model. In the baseline model,

$\widehat{\sigma}^2_{u_{1jk}}$ was 1.72 and statistically significant, whereas in the conditional model, the $\widehat{\sigma}^2_{u_{1jk}}$ is reduced to 0.089. The between-participant variability is small and is not statistically significant ($\widehat{\sigma}^2_{u_{1jk}} = 0.089$, $SE = 0.07$, $Z = 1.35$, $p = 0.09$). Therefore, no additional participant-level moderators will be added to the conditional model. Although the moderator is not significant, it explains a significant amount of between-participant variability. This empirical illustration highlights that coding selected moderators need to be carefully considered as they influence the interpretation of the estimated parameters.

3.4.2 Study and participant moderator

By including age as a participant-level moderator, heterogeneity in effect sizes between studies does not change and remains to be explored. In an attempt to explain variability at the study level, study quality seems to be a promising variable and is added to the model. Study quality is coded as a dummy variable with 0 indicating low-quality studies (i.e., not meeting the WWC design standards) and 1 indicating moderate/high-quality studies (i.e., meeting the WWC design standards with or without reservation). The multilevel meta-analytic model with age as a second-level continuous moderator and quality as a third level dichotomous moderator looks as follows: $b_{1jk} = \gamma_{100} + \gamma_{101} Quality_{101} + \gamma_{200} Age_{11k} + v_{10k} + u_{1jk} + r_{1jk}$. γ_{100} indicates the intervention effect for low-quality studies, and participants at the study average age; γ_{101} reflects the difference between low- and moderate/high-quality studies (controlling for age), and γ_{200} indicates the influence of participant's age on the intervention effectiveness (controlling for study quality). The estimated intervention effect for low-quality studies, and students at the average study age equals 1.74 and remains statistically significant [$\widehat{\gamma}_{100} = 1.74$, $SE = 0.57$, $t(41.7) = 3.06$, $p = 0.0038$]. Controlling for study quality, the influence of age on intervention effectiveness remains 0.14. As anticipated, the higher the quality of the study, the lower the intervention effectiveness (controlling for participant's age). However, this moderator effect is not statistically significant [$\widehat{\gamma}_{101} = -0.27$, $SE = 0.35$, $t(219) = -0.78$, $p = 0.43$]. Therefore, it is not surprising that the between-study variance in intervention effectiveness is not reduced by including quality as a study moderator. It is recommended to explore alternative promising study moderators. Unfortunately, Moeyaert et al. (2019) did not report information about other study-level moderators, and therefore, we could not furtherexplore this. In addition, meta-analysts dependent on information reported by primary study authors. Unfortunately, information related to moderators can be missing in primary SCED studies, or not reported in a useful way. Therefore, Moeyaert et al. (2019) could not code additional moderators. Specific guidelines to report moderators in primary SCED studies can help addressing this issue. For instance, SCED researchers could be encouraged to report specific moderator information by including this as a quality criterion in checklists. SCED primary studies can receive a higher quality ratings if moderator information is reported.

4 Discussion

There is an increased interest in using single-case experimental design studies to evaluate and quantify intervention effectiveness. This results in increased opportunities to summarize intervention effects across studies and help identifying effective evidence-based interventions. The multilevel meta-analytic technique is promising and has been empirically validated (Declercq et al. 2020; Moeyaert et al. 2013a, b; Ugille et al. 2012). However, one complexity that has not been studied is the use of two-stage multilevel meta-analysis to estimate the influence of participant and study-level moderators on intervention effectiveness. This is of crucial importance to make appropriate inferences about intervention effectiveness (for whom is the intervention working, and under which conditions?). This study was designed to provide an introduction to two-stage multilevel meta-analysis, and demonstrate its usefulness to explain intervention heterogeneity by adding moderators. The uniqueness of this model is that the multilayered SCED meta-analytic structure is taken into account and as such moderators at the appropriate participant and study level can be added.

Future methodological research is needed to investigate the statistical properties of the model under a variety of complex design conditions (i.e., non-linear trends, autocorrelation, cross-level interactions, etc.). Additional research is needed to investigate whether there is a limit to the number of participant-level and study-level moderators that can be added to the model, taking the specific small-n characteristics of SCED meta-analyses into account. Further methodological research is needed to investigate the power to estimate intervention and moderator effects, given representative conditions for the field of SCED meta-analyses. Recently, Moeyaert et al. (2021a, b) published a study discussing these conditions and this can be used to design a future Monte Caro simulation study. Moeyaert et al. (2021b) conducted a large-scale Monte Carlo simulation study to investigate the power of the two-level hierarchical linear model to estimate moderators and intervention effects for primary SCED studies. The sizes of the intervention and moderator effects in Moeyaert et al. (2021b) are comparable to the values found in the current study. They found that the more moderators added to the model, the more participants needed to detect the effects of intervention and moderators with sufficient power. If studies include one moderator (nominal with two categories), at least 12 participants are needed to have enough power to capture the intervention effect, while the same studies not only need at least 12 participants but also require a large moderator effect to detect the moderator effect with sufficient power. If including more moderators, at least 20 participants are needed to have sufficient power to detect the intervention and/or moderator effects. The study of Moeyaert et al. (2021b) can be further expanded upon by adding an additional level. A user-friendly tool to pre-process the data, and run the unconditional and conditional two-stage multilevel models is another idea for future research. Xu et al. (2021) developed a user-friendly Shiny tool "PowerSCED" to estimate the power of the two-level model to estimate study-level intervention effect and participant moderators. This tool can be further expanded for meta-analytic purposes.

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

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Alresheed F, Hott BL, Bano C (2013) Single subject research: a synthesis of analytic methods. J Spec Educ Apprenticeship 2(1):1–18

Borenstein M, Hedges LV, Higgins JPT, Rothstein HR (2009a) Introduction to meta-analysis. John Wiley & Sons, Chichester

Borenstein M, Hedges LV, Higgins JPT, Rothstein HR (2009b) A basic introduction to fixed-effect and random-effects models for meta-analysis. Res Synth Methods 1:97–111

Borenstein M, Hedges L, Higgens J, Rohtstein H (2013) Comprehensive meta-analysis (CMA), Version 3. [Computer program]. Biostat, Englewood, NJ

Card NA (2016) Applied meta-analysis for social science research. Guilford

Cooper H (2017) Research synthesis and meta-analysis: a step-by-step approach. Sage Publications, Inc

Declercq L, Jamshidi L, Fernandez Castilla B, Moeyaert M, Beretvas SN, Ferron JM, Van den Noortgate W (2020) Multilevel meta-analysis of individual participant data of single-case experimental designs: one-stage versus two-stage methods. Multivariate Behav Res. https://doi.org/10.1080/00273171.2020.1822148

Ferron JM, Farmer JL, Owens CM (2010) Estimating individual treatment effects from multiple-baseline data: a Monte Carlo study of multilevel-modeling approaches. Behav Res Methods 42(4):930–943

Fingerhut J, Xinyun X, Moeyaert M (2021) Selecting the proper Tau-U measure for single-case experimental designs: development and application of a decision flowchart. Evid Based Commun Interv. https://doi.org/10.1080/17489539.2021.1937851

Glass G (1976) Primary, secondary, and meta-analysis of research. Educ Res 5(10):3–8

Hedges LV, Olkin I (1985) Statistical methods for meta-analysis. Academic Press

Higgins JPT, Li T, Deeks JJ (2021) Chapter 6: Choosing effect measures and computing estimates of effect. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (Eds) Cochrane handbook for systematic reviews of interventions, version 6.2 (updated Feb 2021). Cochrane. Available from www.training.cochrane.org/handbook

Jamshidi L, Heyvaert M, Declercq L, Fernández-Castilla B, Ferron JM, Moeyaert M, Van den Noortgate W (2021) A systematic review of single-case experimental design meta-analyses: characteristics of study designs, data, and analyses. Evid Based Commun Assess Interv

Lipsey MW, Wilson DB (2001) Practical meta-analysis. Sage Publications, Inc

Lobo M, Moeyaert M, Babik I, Cunha A (2017) Single-case experimental design, analysis, and quality assessment for intervention research. J Neurol Phys Ther 14:187–197

Ma H-H (2006) An alternative method for quantitative synthesis of single-subject researches: percentage of data points exceeding the median. Behav Modif 30(5):598–617

Moeyaert M, Ugille M, Ferron J, Beretvas S, Van den Noortgate W (2013a) The three-level synthesis of standardized single-subject experimental data: a Monte Carlo simulation study. Multivar Behav Res 48:719–748

Moeyaert M, Ugille M, Ferron J, Beretvas S, Van Den Noortgate W (2013b) Modeling external events in the three-level analysis of multiple-baseline across-participants designs: a simulation study. Behav Res Methods 45(2):547–559

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. Behav Modif 38(5):665–704

Moeyaert M, Maggin DM, Verkuilen J (2016) Reliability, validity, and usability of data extraction programs for single-case research designs. Behav Modif 40(6):874–900
Moeyaert M, Klingbeil DA, Rodabaugh E, Turan M (2019) Three-level meta-analysis of single-case data regarding the effects of peer tutoring on academic and social-behavioral outcomes for at-risk students and students with disabilities. Remedial Special Educ 42(2):94–106
Moeyaert M, Yang P, Xu X, Kim E (2021a) Characteristics of moderators in meta-analyses of single-case experimental design studies. Behav Modif. https://doi.org/10.1177/01454455211002111
Moeyaert M, Yang P, Xu X (2021b) The power to explain variability in intervention effectiveness in single-case research using hierarchical linear modeling. Perspect Behav Sci
Parker RI, Vannest K (2009) An improved effect size for Single-case research: nonoverlap of all pairs. Behav Ther 40:357–367
Parker RI, Hagan-Burke S, Vannest K (2007) Percentage of all non-overlapping data (PAND): an alternative to PND. J Spec Educ 40(4):194–204
Parker RI, Vannest KJ, Davis JL (2011) Effect size in single-case research: a review of nine nonoverlap techniques. Behav Modif 35(4):303–322
Pustejovsky JE, Chen M, Hamilton B (2021) scdhlm: a web-based calculator for between-case standardized mean differences (Version 0.5.2) \[Web application\]. Retrieved from: https://jepusto.shinyapps.io/scdhlm
Raudenbush SW, Bryk AS (2002) Hierarchical linear models: applications and data analysis methods, 2nd edn. Sage Publications
Review Manager (2020) RevMan (Version 5.4.1) [Computer program]. The Cochrane Collaboration
Saddler B, Asaro-Saddler K, Moeyaert M, Ellis-Robinson T (2017) Effects of a summarizing strategy on written summaries of children with emotional and behavioral disorders. Remed Spec Educ 38(2):87–97
SAS Institute Inc (2014) SAS software (Version 9.4). Retrieved from https://sas.com
Scruggs TE, Mastropieri MA, Casto G (1987) The quantitative synthesis of single-subject research: methodology and validation. Remed Spec Educ 8(2):24–33
Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F (2000) Methods for meta-analysis in medical research. John Wiley & Sons, Chichester
Ugille M, Moeyaert M, Beretvas SN, Ferron J, Noortgate W (2012) Multilevel meta-analysis of single-subject exper- imental designs: a simulation study. Behav Res Methods 44:1244–1254
Van den Noortgate W, Onghena P (2003a) Combining single-case experimental data using hierarchical linear models. Sch Psychol Q 18(3):325–346
Van den Noortgate W, Onghena P (2003b) Hierarchical linear models for the quantitative integration of effect sizes in single-case research. Behav Res Methods Instrum Comput 35(1):1–10
Van den Noortgate W, Onghena P (2007) The aggregation of single-case results using hierarchical linear models. Behav Analyst Today 8:196–209
Van den Noortgate W, Onghena P (2008) A multilevel meta-analysis of single-subject experimental design studies. Evid Based Commun Assess Interv 2(3):142–151
Van den Noortgate W, Opdenakker M, Onghena P (2005) The effects of ignoring a level in multilevel analysis. Sch Eff Sch Improv 16(3):281–303
What Works Clearinghouse (2020) What Works Clearinghouse™ Procedures Handbook (Version 4.1). Retrieved from https://ies.ed.gov/ncee/wwc/Handbooks
Wilson DB (n.d.) Practical meta-analysis effect size calculator [Online calculator]. Retrieved Month Day, Year, from https://campbellcollaboration.org/research-resources/effect-size-calculator.html
Xu X, Moeyaert M, Yang P (2021) PowerSCED (Version 1.0) [Web application]. Retrieved from https://xinyunxu.shinyapps.io/PowerSCED/_w_8b4d5ac0/
Zelinsky NA, Shadish W (2018) A demonstration of how to do a meta-analysis that combines single-case designs with between-groups experiments: the effects of choice making on challenging behaviors performed by people with disabilities. Dev Neurorehabil 21(4):266–278

Authors and Affiliations

Mariola Moeyaert[1] · **Panpan Yang**[2]

Panpan Yang
py5452@princeton.edu

1 School of Education, Department of Educational and Counseling Psychology, Division of Educational Psychology & Methodology, The University at Albany - State University of New York, 1400 Washington Ave, Albany, NY 12222, USA

2 Department of Population Research, Princeton University, Princeton, NJ, USA