Hammill Institute on Disabilities

Behavioral Disorders
2019, Vol. 44(4) 241–256
© Hammill Institute on Disabilities 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0198742918806926
journals.sagepub.com/home/bhd
SAGE

Quantitative Synthesis of Research Evidence: Multilevel Meta-Analysis

Mariola Moeyaert[1]

Abstract

Multilevel meta-analysis is an innovative synthesis technique used for the quantitative integration of effect size estimates across participants and across studies. The quantitative summary allows for objective, evidence-based, and informed decisions in research, practice, and policy. Based on previous methodological work, the technique results in powerful, unbiased, and precise effect size estimates. However, its use in practice is limited and its full potential is not yet fully understood. This article aims to bring the multilevel meta-analytic model closer to the applied researcher by introducing the technique at a conceptual level and discussing its full potential and relevance to the field. The procedure of multilevel meta-analysis is illustrated using a recent single-case meta-analytic dataset. Software codes, output tables, interpretations, and graphical displays of effect size estimates are given such that the reader can repeat the analysis independently.

Keywords

multilevel meta-analysis, regression-based effect size, precision, standardization, Hedges's bias correction, single-case experimental design

Multilevel meta-analysis is a promising analysis technique used to quantitatively summarize research findings across similarly focused studies (Van den Noortgate & Onghena, 2007). Given the increased number of published studies investigating related underlying research questions, and the importance of replication studies, the use of multilevel meta-analysis becomes increasingly important. Multilevel meta-analysis is a valuable means to inform unbiased evidence-based decisions that are valid and reliable (Kratochwill et al., 2010; Ugille, Moeyaert, Beretvas, Ferron, & Noortgate, 2012). For instance, in the field of education, a practitioner might be interested in whether peer-tutoring interventions are effective to improve social skills (e.g., number of peer interactions) for students with behavior disorders. The quantitative synthesis of all studies investigating this research question can provide meaningful estimates of the intervention's anticipated effect on social behaviors. It would be unfortunate to ignore research investments and evidence that is already available in the literature given that high-quality meta-analyses can result in important insights for policy makers, funding agencies, practitioners in the field, and researchers (Talbott, Maggin, Van Acker, & Kumm, 2018).

The multilevel meta-analytic model is particularly useful for summarizing hierarchical structured data such as single-case experimental design studies (SCEDs; Van den Noortgate & Onghena, 2003a, 2003b). In SCEDs, the effectiveness of a treatment is usually evaluated across multiple participants resulting in multiple dependent effect sizes per studies. For example, in the study of Mason et al. (2014), the number of communicative acts for three students with autism spectrum disorders was measured repeatedly during a baseline condition before introducing the peer-tutoring intervention (i.e., treatment condition). Therefore, three estimates of the effectiveness of peer-tutoring on communicative acts are obtained (see Figure 1). Traditional meta-analyses synthesize study-specific effect sizes across studies, whereas multilevel meta-analyses are capable of summarizing participant-specific effect sizes across cases and across studies. Therefore, the multilevel meta-analytic model is needed to summarize evidence originating from SCEDs.

The statistical properties of the multilevel meta-analysis method have been intensively studied and validated during the last decade using large-scale Monte Carlo simulation studies (Moeyaert, Ugille, Ferron, Beretvas, & Van den Noortgate, 2013, 2014b; Ugille et al., 2012). However, it is used infrequently in practice, perhaps because (a) the analysis may appear overly complex and/or (b) the technique is relatively new to the field of education and its potentials

[1] University at Albany, NY, USA

Corresponding Author:
Mariola Moeyaert, Division of Educational Psychology & Methodology, Department of Educational and Counseling Psychology, School of Education, The State University of New York, University at Albany, 1400 Washington Ave., Albany, NY 12222, USA.
Email: mmoeyaert@albany.edu

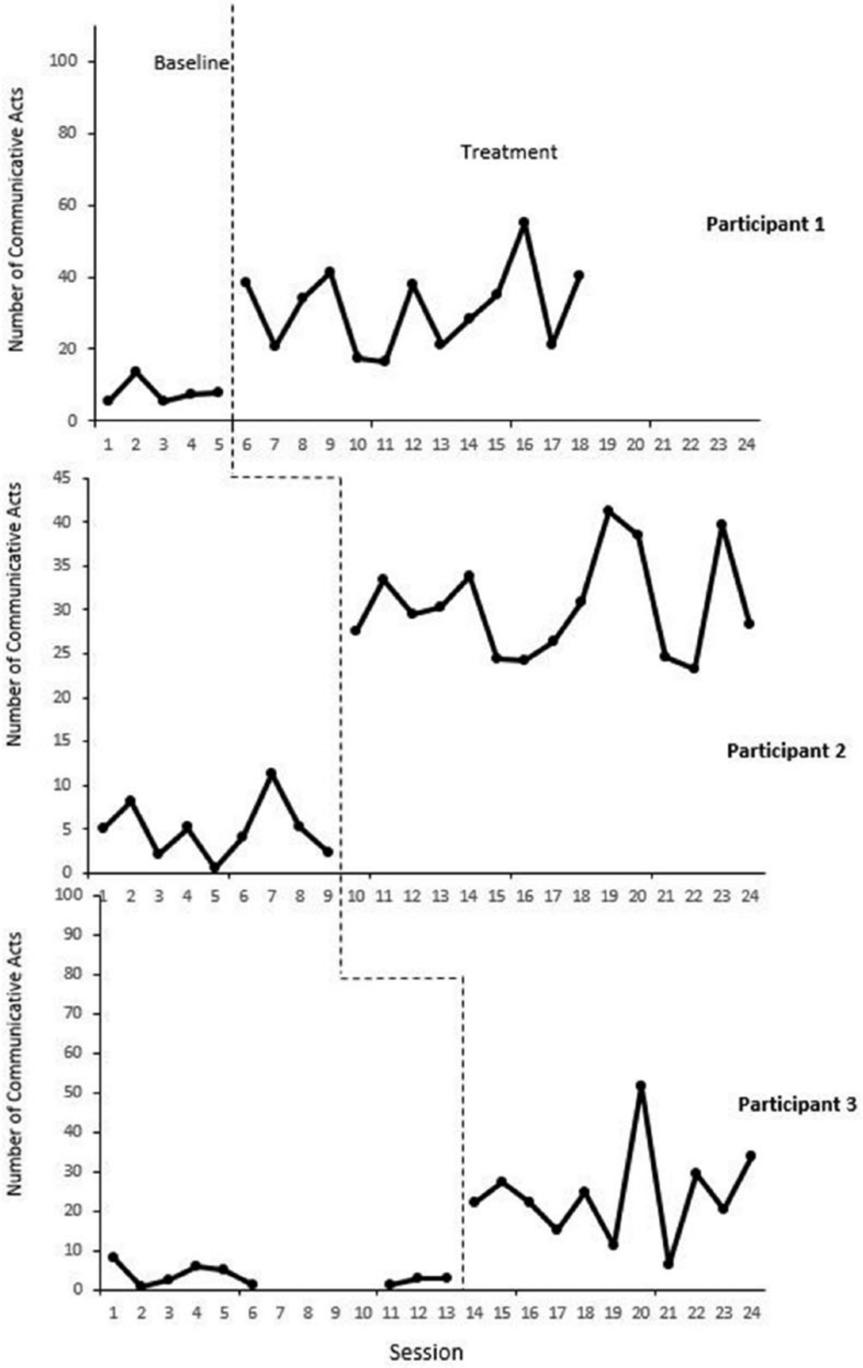

Figure 1. Graphical display of a single-case experimental design study in which the effectiveness of a treatment is evaluated across three participants.
Note. Raw data for the graphical presentation were retrieved from graphs presented in Mason et al. (2014).

and applications are not yet widely understood. Indeed, a recent systematic review of meta-analyses of SCEDs (including a total of 178 studies) indicated that only a small percentages (i.e., 17%) of the meta-analyses used multilevel meta-analysis (Jamshidi et al., 2018). Therefore, this article is a first attempt to (a) introduce to multilevel meta-analytic model to applied researchers, (b) give a basic conceptual understanding of the multilevel meta-analytic procedure, and (c) enhance the use of the technique by giving a step-by-step demonstration of the procedure applied to a real meta-analytic dataset and providing documented statistical software code (SAS 9.4, Copyright © 2017, SAS Institute Inc., SAS). The aim is to enhance the field as a whole by introducing high-quality synthesis techniques.

Multilevel Meta-Analysis: General Introduction

Similar to traditional meta-analyses, multilevel meta-analysis is a practical and useful tool for systematically evaluating research evidence across primary studies investigating the same underlying research question (Glass, 1976). Gene Glass is known as the founding father of meta-analysis and introduced this technique to the field of social sciences. In 1977, he published an article together with Smith summarizing the effect of psychotherapy using meta-analysis (Smith & Glass, 1977). In general, five steps can be followed to successfully conduct a (multilevel) meta-analysis. First, start with formulating a research question of interested. This will inform the inclusion criteria (population, outcomes, interventions, variables of interest, etc.). Second, relevant literature is searched (ideally by multiple independent researchers) including publications, dissertations, technical report, theses, and so on. Third, the data are retrieved and variables are coded (for post hoc calculation of effect sizes). In this step, it is important to gather as much detail as possible. Fourth, once all the data are coded, the multilevel meta-analysis can be conducted. Fifth, the results are reported, interpreted, and discussed. The focus in the current article is giving guidance on Steps 4 and 5, as the previous steps are identical to traditional meta-analyses and systematic reviews in general. Details about the traditional meta-analytic procedure can be found in Borenstein, Hedges, Higgins, and Rothstein (2009); Card (2012); Hedges and Olkin (1985); Lipsey and Wilson (2001); and Sutton, Abrams, Jones, Sheldon, and Song (2000).

The research evidence reported in primary studies can be summarized by an effect size measure. There are two primary classes of effect size measures (Lipsey & Wilson, 2001), namely, the standardized mean difference—representing the size and direction of the difference between two groups' sample means expressed in standard deviations, $\delta((\mu_E - \mu_C)/\sigma)$, and the correlation coefficient—representing the strength and direction of the relationship between two variables, ρ. Recently, methods have been derived and validated to summarize other kinds of effect sizes, such as regression coefficients, proportions, odds ratios, standard deviations, and reliability coefficients. Depending on your research question (and the primary-level data that are available), you can choose the most relevant effect size. Alternatively, if the summary statistic of interest is not reported in the primary study, it can be calculated using the raw data reported in the primary study (or extracted from the graphical presentation of the raw data, which is very common in SCEDs; see below). If a summary statistic other than the one of interest is reported, several conversion formulae can be applied. For instance, if you are interested in summarizing Hedges's *g* (a common standardized mean difference effect size) but only the correlation coefficient is given, then Hedges's *g* can be easily converted from correlation coefficient using the following formula: $g = 2r/\sqrt{1-r^2}$. For an in-depth discussion and other conversion formulae, see Lipsey and Wilson (2001) and Rosenthal (1994).

Meta-analysis is also called the analysis of analyses, and results in a single best estimate (usually a weighted average) of the effect size of interest. There are several convincing reasons to consider meta-analyzing primary studies' effect sizes, including (a) generalizing research findings, (b) identifying areas where more research is needed, (c) dealing with subjectivity of verbal narrative literature reviews, (d) enhancing power for statistical tests (i.e., larger sample sizes result in more precise estimates), and as with all statistics (e) parsimony. Because we are pooling effect sizes together from several studies, a more precise effect size estimate is obtained (i.e., smaller standard error [*SE*]) compared with one single effect size estimate. As a consequence, we can be more confident in generalizing the research findings. In addition, variability in effect size estimates between studies can be explored by including moderators (e.g., the effectiveness of a treatment might depend on gender, or the relation between depression and anxiety might be explained by age).

As an example, Raudenbush and Bryk (1985) meta-analyzed 19 studies investigating how teachers' expectations about their students can influence the actual IQ. The standardized mean difference was the effect size of interest and calculated per primary study. Table 1 contains the standardized mean difference effect sizes per study (Y_i) with their corresponding *SE*. For instance, Y_3 refers to the standardized mean difference for Study 3 (i.e., Jose & Cody, 1971) and equals −0.14 ($SE = 0.16$) whereas Y_4 is 1.18 ($SE = 0.37$; Pellegrini & Hicks, 1972). This illustrates the existence of variability in the size, direction, and precision of the estimated relation between teacher's expectations and IQ. The effect size for Study 4 is positive and larger in magnitude but less precise compared with the effect size for Study 3. Consequently, there is inconsistency in evidence. Depending on the study examined, different inferences will be made regarding the intervention's effectiveness. Therefore, instead of relying on just one study to draw inferential conclusions, the effect sizes of the 19 primary studies can be pooled together, weighted by the inverse of the squared *SE* (i.e., studies with a lower *SE* are more precise and, as a consequence, are given more weight in the meta-analysis). In

Table 1. Meta-Analytic Dataset Raudenbush and Bryk (1985).

Study no.	Study	Weeks	Y_i	*SE*	Var	Precision
1	Rosenthal et al. (1974)	2	0.0300	0.1249	0.0156	64.1026
2	Conn et al. (1968)	21	0.1200	0.1470	0.0216	46.2963
3	Jose and Cody (1971)	19	−0.1400	0.1670	0.0279	35.8423
4	Pellegrini and Hicks (1972)	0	1.1800	0.3730	0.1391	7.1891
5	Pellegrini and Hicks (1972)	0	0.2600	0.3691	0.1362	7.3421
6	Evans and Rosenthal (1969)	3	−0.0600	0.1030	0.0106	94.3396
7	Fielder et al. (1971)	17	−0.0200	0.1030	0.0106	94.3396
8	Claiborn (1969)	24	−0.3200	0.2200	0.0484	20.6612
9	Kester (1969)	0	0.2700	0.1640	0.0269	37.1747
10	Maxwell (1970)	1	0.8000	0.2510	0.0630	15.8730
11	Carter (1970)	0	0.5100	0.3020	0.0912	10.9649
12	Flowers (1966)	0	0.1800	0.2229	0.0497	20.1207
13	Keshock (1970)	1	−0.0200	0.2890	0.0835	11.9760
14	Henrikson (1970)	2	0.2300	0.2900	0.0841	11.8906
15	Fine (1972)	17	−0.1800	0.1591	0.0253	39.5257
16	Grieger (1970)	5	−0.0600	0.1670	0.0279	35.8423
17	Rosenthal and Jacobson (1968)	1	0.3000	0.1389	0.0193	51.8135
18	Fleming and Anttonen (1971)	2	0.0700	0.0938	0.0088	113.6364
19	Ginsburg (1970)	7	−0.0700	0.1741	0.0303	33.0033

Note. Y_i indicates the standardized mean difference, *SE* indicates the standard error, Var indicates the variance, and Precision is the inverse of the variance.
For complete reference information for these studies, see Raudenbush & Bryk, 1985.

addition, sources of variability between the effects sizes can be explored.

The overall weighted average effect size estimate across the 19 studies equals 0.084 ($SE = 0.052$, $Z = 1.621$, $p = .105$). This means that the higher the teachers' expectations, the higher the actual students IQ levels. The result is not statistically significant (two-tailed testing, $\alpha = .05$) and a significant amount of between-study variability is found (see Raudenbush & Bryk, 1985, for more details). Therefore, in a next step, a moderator was added because it can be expected that the variability in the relation between teachers' expectations and student's actual IQ scores can be partially explained by the number of prior weeks of contact. Indeed, if "Weeks" (see Table 1) was added as moderator, the standardized mean difference equaled 0.407 ($SE = 0.087$, $Z = 4.678$, $p < .001$) and the effect of "weeks" equaled −0.157 ($SE = 0.036$, $Z = −4.388$, $p < .001$). This means that there is a statistically significant positive relation between teachers' expectations and actual IQ, but the more the prior contact, the smaller this relationship becomes.

For the meta-analysis of SCEDs, the regression coefficient will be used as the effect size given the nature of the single-case design (i.e., repeated measures over time during control and treatment sessions, see Figure 1). The regression-based effect size is recommended in this context as it can account for data trends, between-phase variability, and autocorrelation, and has a known sampling distribution (Kratochwill et al., 2010; Lenz, 2013; Parker & Vannest, 2008; Shadish, Rindskopf, Hedges, & Sullivan, 2012). In SCEDs, the repeated measures across time are graphically presented as demonstrated in Figure 1, and as a consequence, the raw data can be obtained by using a data retrieval software programs such as WebPlotDigitizer, Datathief, XYit, and Ungraph (Moeyaert, Maggin, & Verkuilen, 2016). These data retrieval programs are user-friendly, point-and-click software. This allows for calculating the regression-based effect size and *SE*. As mentioned before and illustrated in Figure 1, in the area of SCEDs, the effectiveness of an intervention is usually replicated across participants resulting in multiple effect size measures per study. This is usually not the case in a group-comparison design study in which one standardized mean difference between the experimental and control condition is reported. As such, effect sizes in SCEDs within one study are dependent. If we simply combine effects across cases and ignore the study level, we assume that we have more information available than there is in reality. As a consequence, the effect sizes are estimated more precisely, resulting in *SE* estimates that are too small. Smaller *SE*s result in larger test statistics (and smaller *p* values), and therefore, it becomes more easily to reject a null hypothesis (increasing the likelihood of making Type I errors or falsely rejecting a true null hypothesis). Therefore, in contexts of studies containing more than one effect size per study, multilevel meta-analysis (as opposed to a traditional meta-analysis) is most appropriate and recommended for the quantitative synthesis.

Multilevel Meta-Analysis: Methodology and Empirical Illustration

When conducting a multilevel meta-analysis of SCEDs, the first step is to calculate participant-specific standardized effect sizes. I will start this section discussing how standardized regression effect sizes can be estimated. This involves multiple steps: (a) obtaining raw SCED data, (b) running a single-level regression model per participant, (c) standardizing the regression coefficients, and (d) correcting the standardized regression coefficients for small-sample bias. In the second part of this section, I will demonstrate how the standardized and bias-corrected effect size estimates can be combined using the multilevel meta-analysis model.

Single-Level Analysis: Standardized Regression-Based Effect Size

Raw data extraction. I illustrate the procedure for estimating the regression-based effect size using the study of Mason et al. (2014), displayed in Figure 1, which is part of a bigger meta-analytic dataset summarizing the effects of peer-tutoring interventions on academic and social outcome scores (Moeyaert, Klingbeil, Rodabaugh, & Turan, 2018). The first step involves raw SCED data extraction from the primary study graphs. This is necessary, as it is unlikely that the regression-based effect size together with its *SE* is reported in the primary studies. Indeed, Jamshidi et al. (2018) found that only a small percentage (i.e., 1.90%) of SCEDs published between 1985 and 2015 used and reported regression-based effect sizes. The free data retrieval software program WedPlotDigitizer (Rohatgi, 2014) was used for this purpose (and can be downloaded for free: https://automeris.io/WebPlotDigitizer/). The graph image of the primary study can be imported in the data extraction program to get the *X* values (e.g., session numbers in Figure 1) and the *Y* values (e.g., number of communicative acts per 10-min session in Figure 1). For this purpose, the axes need to be calibrated (e.g., you "tell" the program where the axis is and what the minimum and maximum values are). Then, you individually select each data point by "clicking" on it with a mouse.

Effect size calculation. The raw data can then be used to estimate the treatment effect(s) per participant (i.e., regression-based effect size estimate). For this purpose, an ordinary least squares (OLS) regression model was built. For instance, a researcher might be interested in quantifying a change in outcome score between the last measure of the baseline and the first measure of the treatment phase (i.e., immediate treatment effect) and a change in linear trend between the baseline and the treatment phase (i.e., treatment effect on the time trend):

$$\begin{aligned} y_{ijk} &= \beta_{0jk} + \beta_{1jk}\mathrm{Time}_{ijk} + \beta_{2jk}\mathrm{Treatment}_{ijk} \\ &\quad + \beta_{3jk}\mathrm{Time}'_{ijk} \times \mathrm{Treatment}_{ijk} + e_{ijk}, \end{aligned} \qquad (1)$$

where i stands for the measurement occasion ($i = 0,1,\ldots I$), j for the case ($j = 1,2,\ldots J$), and k for the study ($k = 1,2,\ldots K$). This means that y_{ijk} indicates the outcome score on measurement occasion i for case j from study k. Before running the OLS regression, a design matrix is created. A design matrix contains all exploratory variables of interest (e.g., Time_{ijk}, $\mathrm{Treatment}_{ijk}$, and $\mathrm{Time}'_{ijk} \times \mathrm{Treatment}_{ijk}$) and values are assigned to each variable. An example of a possible design matrix for the study of Mason et al. (2014) is given in Table 2. Table 2 contains three exploratory variables: $\mathrm{Treatment}_{ijk}$ is a dummy coded variable indicating whether the measurement occasion belongs to the baseline phase ($\mathrm{Treatment}_{ijk} = 0$) or the treatment phase ($\mathrm{Treatment}_{ijk} = 1$); Time_{ijk} is a time-related variable that equals 0 on the first measurement occasion of the baseline phase and increases by one unit for all subsequent measurement occasions. In addition, an interaction variable is created between the centered time-indicator (Time'_{ijk}) and the dummy variable ($\mathrm{Treatment}_{ijk}$). Time'_{ijk} is centered such that it equals 0 on the first measurement occasion of the treatment phase (i.e., $\mathrm{Time}'_{ijk} = [\mathrm{Time}_{ijk} - (n_{Aj} + 1)]$). Careful attention needs to be paid to set up the design matrix accordingly because depending on the coding of the design matrix, different effect sizes can be calculated (Moeyaert, Ugille, Ferron, Beretvas, & Van den Noortgate, 2014a).

Table 2 gives a display of the raw data and design matrix for Participant 2 from the Mason et al. study. Note that the same coding matrix needs to be created for all the participants included in the meta-analysis. By setting the design matrix up this way, the following regression effect sizes are obtained (and can be used afterward for quantitative synthesis): β_{0jk} indicating the expected baseline level, β_{1jk} equaling the expected linear trend during the baseline, β_{2jk} referring to the expected immediate treatment effect, and β_{3jk} referring to the expected effect of the treatment on the time trend (i.e., a change in slope). By applying Equation 1 to the raw data of Participant 2, four estimated effect sizes are obtained: (a) The estimated initial baseline level at the start of the baseline, b_{0jk}, $b_{0jk} = 5.092$, $SE = 3.25$, $t(20) = 1.56$, $p = .13$; (b) the linear trend during the baseline phase, which is estimated to be slightly negative, $b_{1jk} = -0.049$, $SE = 0.683$, $t(20) = -0.071$, $p = .944$; (c) the estimated immediate treatment effect, b_{2jk}, estimated to be large, positive, and statistically significant, $b_{2jk} = 22.855$, $SE = 6.166$, $t(20) = 3.707$, $p = .001$; and (d) the change in trend between baseline phase and treatment phase, b_{3jk}, $b_{3jk} = 0.201$, $SE = 0.201$, $t(20) = 0.268$, $p = .792$. The results indicate that the peer-tutoring intervention results in an immediate increase in social skills for Participant

Table 2. Raw Data and Coding of the Design Matrix of Participant 1 From Mason et al. (2014).

Participant	Session	Y (outcome)	Time	Treatment	Time′ × Treatment	Time_3′	Time_3′ × Treatment
1	1	5.28	0	0	−6	−8	0
1	2	13.38	1	0	−5	−7	0
1	3	5.27	2	0	−4	−6	0
1	4	7.42	3	0	−3	−5	0
1	5	7.75	4	0	−2	−4	0
1	6	7.41	5	0	−1	−3	0
1	7	38.15	6	1	0	−2	0
1	8	20.46	7	1	1	−1	0
1	9	34.18	8	1	2	0	0
1	10	41.45	9	1	3	1	1
1	11	17.31	10	1	4	2	2
1	12	16.31	11	1	5	3	3
1	13	38.12	12	1	6	4	4
1	14	21.25	13	1	7	5	5
1	15	28.19	14	1	8	6	6
1	16	35.12	15	1	9	7	7
1	17	55.10	16	1	10	8	8
1	18	21.20	17	1	11	9	9
1	19	40.19	18	1	12	10	10

2 (i.e., b_{2jk} is statistically significant). The regression coefficients of interest that will be used in the multilevel meta-analysis (i.e., b_{2jk} and b_{3jk}) are presented in Figure 2 for Participant 2 (i.e., Ed). The user-friendly web application MultiSCED (Cools et al., 2017) was used to graphically present the estimated OLS regression lines per participant of the study of Mason et al. (2014; see Figure 3). The MultiSCED tool (http://52.14.146.253/MultiSCED/) is an open source environment and comes with a user guide. The user guide provides a step-by-step OLS analysis resulting in the regression-based effect sizes of interest. The meta-analytic dataset of Shogren, Fagella-Luby, Bae, and Wehmeyer (2004), which is freely available (https://kuleuven.app.box.com/v/Shogren2004), is used to demonstrate the two-level and three-level meta-analysis to combine regression-based effect size estimates. All the steps are supplemented with print screens of the environment.

Centering of the time variable. Depending on the researcher's interest, the time variable can be centered around another measurement point in the treatment phase. For instance, if the researcher wants to evaluate the effectiveness of the treatment at the third point in the intervention, then the time variable of the interaction term (Time_3′) should be centered around that value. This is represented by the variable Time_3′ × Treatment in Table 2. The same regression output will be obtained as in Equation 1, except from the estimated treatment effect (i.e., b_{2jk}) as this is now represented as the difference between the predicted outcome score during the third intervention session and the predicted baseline score at that point, $b_{2jk} = 24.668$, $SE = 4.64$, $t(20) = 5.316$, $p < .001$. This is visualized in Figure 4. This demonstrates that the regression-based effect size estimate is very flexible and allows the effectiveness of the treatment to be estimated at different moments during the intervention phase. The research interest in this article lies in the immediate treatment effect, and therefore, the time variable (labeled as session in Figures 2 and 3) in the interaction term is centered on the first measurement occasion of the treatment. More information about coding the design matrix and the influence of centering time on the interpretation of the obtained treatment effect estimates can be found in Moeyaert, Ugille, Ferron, Beretvas, & Van den Noortgate (2014a). The OLS regression model presented in Equation 1 can easily be extended, reflecting more complex SCED designs (i.e., reversal designs and alternating treatment designs) and SCED data characteristics (i.e., nonlinear trends, autocorrelation), but this is beyond the scope of this study.

Standardization and bias correction. The underlying research interest lies in making statistical inferences regarding the effect size estimate(s). In this context, we want to evaluate whether the treatment effect estimates b_{2jk} and b_{3jk} are statistically significant. In other words, does the peer-tutoring intervention result in an immediate increase in positive social behavior and how is this changing over time (i.e., is the intervention becoming more effective or less effective over time)? Therefore, the primary study

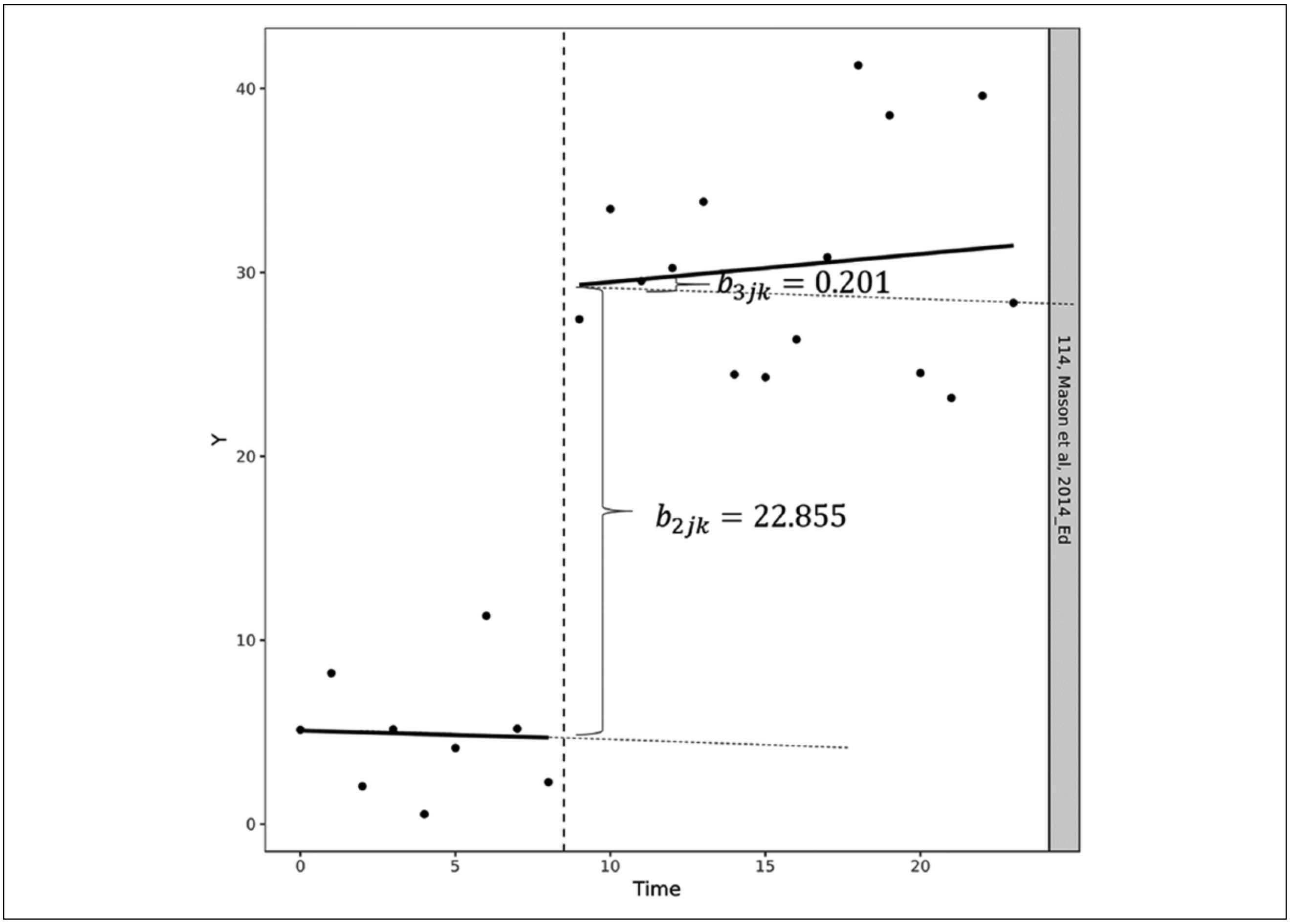

Figure 2. Graphical representation of the immediate treatment effect estimate (b_{2jk}) and the treatment effect on the time trend estimate (b_{3jk}) for Participant 2 (i.e., Ed) from the Mason et al. (2014) study.

summary statistics (i.e., b_{2jk} and b_{3jk}) will be combined across participants and across studies to make generalizations to the broader population. Prior to combining data from different studies, the scores need to be standardized, as it is likely that outcome variables from different studies (and even from different participants within the same study) are measured on different scales. Standardized regression coefficients are obtained by dividing the estimated regression coefficient by the estimated within-case residual standard deviation (i.e., $\hat{\sigma}_{ejk}$ or root mean square error [RMSE]) that is obtained by running the OLS regression as presented in Equation 1. The formulas that can be used to standardize effect sizes are

$$b'_{2jk} = \frac{b_{2jk}}{\hat{\sigma}_{ejk}} \text{ and } b'_{3jk} = \frac{b_{3jk}}{\hat{\sigma}_{ejk}}. \qquad (2)$$

Consequently, the sampling error variance should be divided by the estimated residual error variance:

$$\sigma^{2'}_{r2jk} = \frac{\sigma^{2}_{r2jk}}{\hat{\sigma}^{2}_{ejk}} \text{ and } \sigma^{2'}_{r3jk} = \frac{\sigma^{2}_{r3jk}}{\hat{\sigma}^{2}_{ejk}}, \qquad (3)$$

where $\sigma^{2'}_{r2jk}$ and $\sigma^{2'}_{r3jk}$ indicate the standardized squared *SE*s of b_{2jk} and b_{3jk} , respectively. More details about the standardization formula for regression-based effect sizes can be found in Van den Noortgate and Onghena (2008) and Ugille et al. (2012). The meta-analytic dataset that will be used for quantitative synthesis is compromised of the estimated standardized effect sizes and associated standardized sampling error variance.

There is one extra step to perform specific for combining SCED regression effect sizes. Previous methodological work (Ugille et al., 2012) has indicated that standardization induces some bias when a small number of measurements within a participant are obtained, which is usually the case in contexts of SCEDs (Shadish & Sullivan, 2011). One way to deal with this is correcting the standardized effect sizes (i.e., b'_{jk} in Equation 2) for small-sample bias by

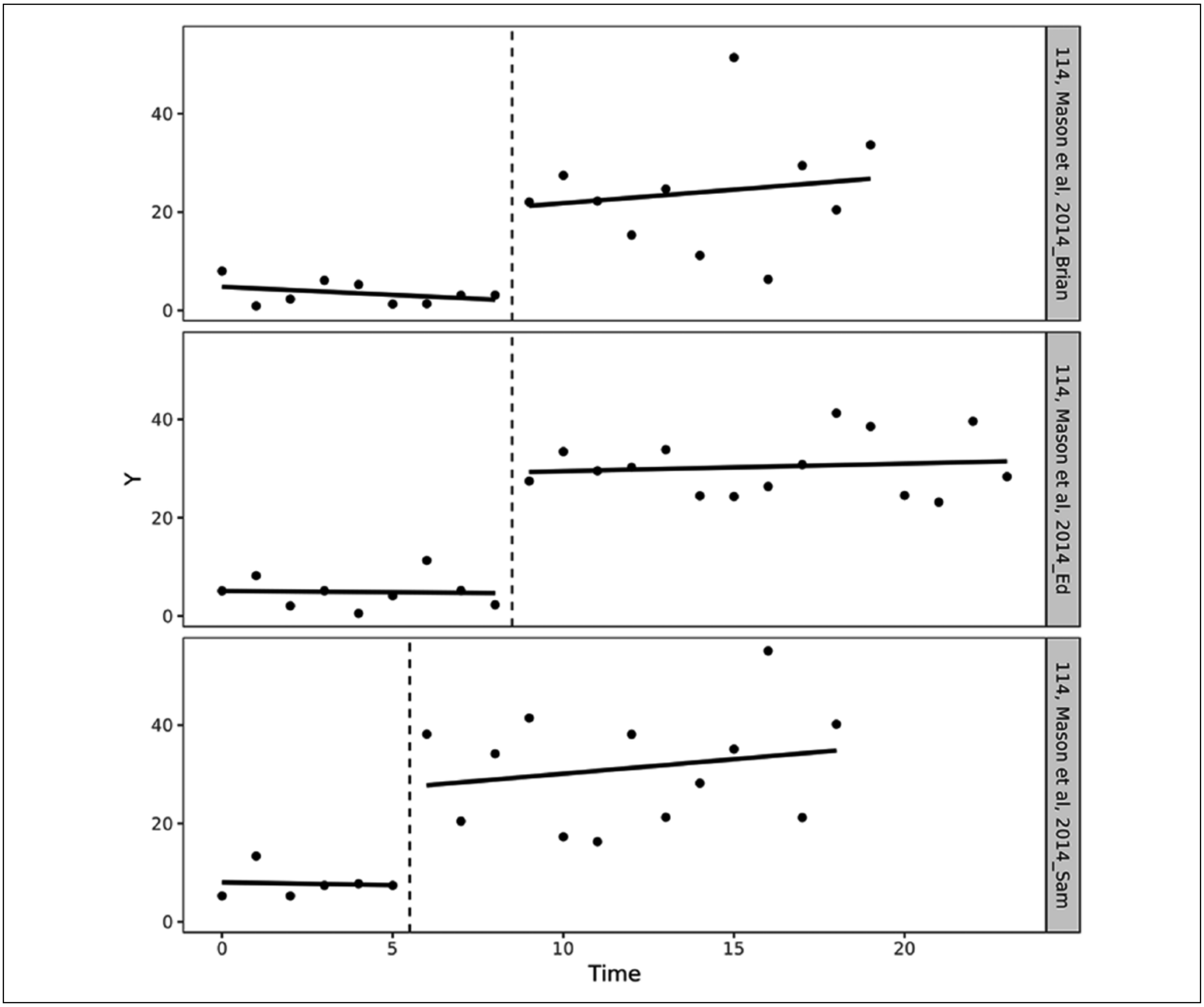

Figure 3. Graphical display of the estimated ordinary least square regression lines for the participants of Mason et al. (2014) using the MultiSCED environment developed by Declercq et al. (2017).

multiplying the standardized effect size by Hedges's bias correction factor (Hedges, 1981; Ugille, Moeyaert, Beretvas, Ferron, & Van den Noortgate, 2013), which is approximately equal to 1 – [3 / (4*m* – 1)], with *m* indicating the degrees of freedom. In the model discussed in Equation 1, *m* equals the number of measurement occasions (*I*) minus the number of predictors (*p*) in the regression model minus 1 (i.e., *m* = *I* – *p* – 1). The corrected standardized immediate treatment effect equals

$$b'^{C}_{2jk} = b'_{2jk}\left(1-\frac{3}{4\left(I-p-1\right)-1}\right) \text{ and }$$
$$b'^{C}_{3jk} = b'_{3jk}\left(1-\frac{3}{4\left(I-p-1\right)-1}\right). \qquad (4)$$

Consequently, the sampling error variance should also be corrected for small bias:

$$\sigma^{2'c}_{r2jk} = \sigma^{2'}_{r2jk}\left(1-\frac{3}{4(I-p-1)-1}\right)^{2} \text{ and }$$
$$\sigma^{2'c}_{r3jk} = \sigma^{2'}_{r3jk}\left(1-\frac{3}{4(I-p-1)-1}\right)^{2}. \qquad (5)$$

The full meta-analytic dataset containing the regression-based effect size estimates, the standardized regression-based effect size estimates, and the bias-corrected standardized effect size estimates together with the appropriate sampling variance (i.e., standardized, bias-corrected) can be found in the supplementary materials.

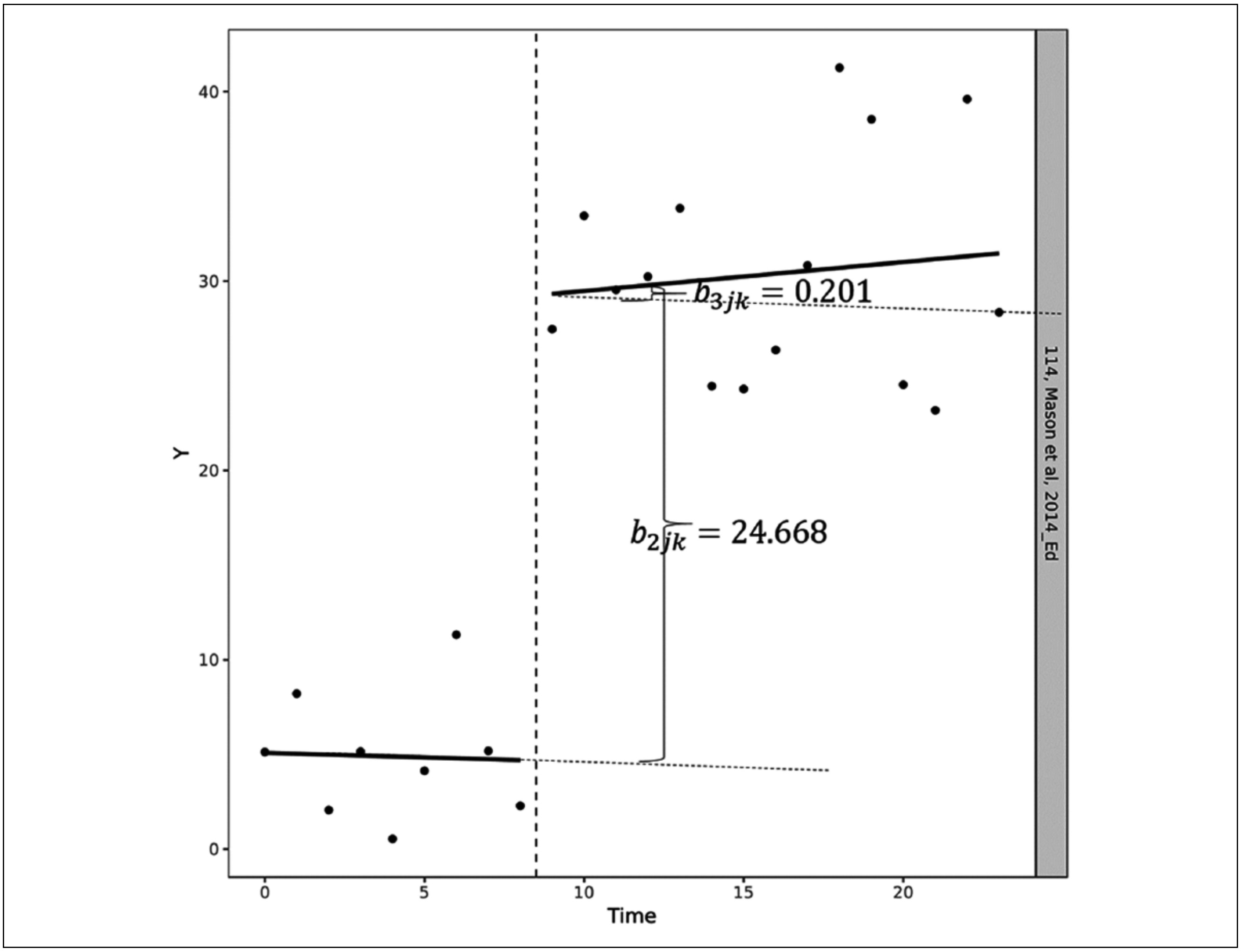

Figure 4. Graphical representation of the effect size estimates of the treatment effect at the third session during the intervention phase (b_{2jk}) and the treatment effect on the time trend (b_{3jk}) for Participant 2 (i.e., Ed) from Mason et al. (2014).

Multilevel Meta-Analysis: Combining Standardized Regression-Based Effect Sizes

The meta-analytic dataset used for the demonstration of the multilevel meta-analysis procedure is from Moeyaert et al. (2018). Moeyaert et al. coded data from 65 SCEDs investigating peer-tutoring as an intervention to increase academic and social outcome scores. We will focus on the studies investigating social outcomes (27 SCEDs, with a total of 130 cases). An overview of the studies included in the multilevel meta-analysis is included in Supplemental Appendix A. SAS Proc Mixed within SAS 9.4 (Copyright © 2017, SAS Institute Inc., SAS) was used to perform the multilevel meta-analysis. The SAS code together with step-by-step descriptions and output tables can be found in Supplemental Appendix B. The Kenward–Roger method (Kenward & Roger, 1997) for estimating degrees of freedom was chosen as it contains a small-sample bias correction that is recommended in single-case contexts (Ferron, Bell, Hess, Rendina-Gobioff, & Hibbard, 2009). Ferron et al. (2009) conducted a large-scale Monte Carlo simulation study comparing five different methods to estimate the degrees of freedom (i.e., residual, containment, between-within, Satterthwaite, and Kenward–Roger) in context of multilevel modeling of SCED data and found that the Kenward–Roger method resulted in the least biased *SE*s estimates of the regression coefficients and (co)variance components. The Kenward–Roger method to estimate the degrees of freedom is described in detail elsewhere (Schaalje, McBride, & Fellingham, 2001).

As mentioned before, first standardized bias-corrected effect sizes and *SE*s (i.e., root square of the standardized bias-corrected sampling error variance) are estimated per participant and per study. In a next step, the estimated effect sizes of the immediate treatment effect, $b_{2jk}^{\prime c}$, and the treatment effect on the time trend, $b_{3jk}^{\prime c}$, for participant *j* from study k are modeled as a function of the average effects sizes, β_{2jk} and β_{3jk}, respectively, plus random deviations, r_{2jk} and r_{3jk}, that are assumed to be normally distributed with a mean of zero. Level 1 of the three-level meta-analytic model looks as follows:

$$\begin{aligned} b'^{c}_{2jk} &= \beta_{2jk} + r_{2jk} \text{ and } r_{2jk} \quad N\left(0, \sigma^{2'c}_{r2jk}\right), \\ b'^{c}_{3jk} &= \beta_{3jk} + r_{3jk} \text{ and } r_{3jk} \quad N\left(0, \sigma^{2'c}_{r3jk}\right). \end{aligned} \tag{6}$$

The sampling error variances of the observed effects, $\sigma^2_{r_{2jk}}$ and $\sigma^2_{r_{3jk}}$, are typically reported by default for each estimated OLS regression coefficient when performing a regression analysis (and we can standardize and bias correct them as explained above). In a meta-analysis (and in the multilevel meta-analysis), these variances are treated as "known" (for more information about this, see Lipsey & Wilson, 2001). These variances depend, to a large extent, on the number of observations and the variance of these observations, and therefore can be participant- and study-specific. It makes sense to weight effect sizes by their precision, assigning more weight to effect sizes that are more precise. Precision is defined as the inverse of the sampling error variances: $1/\sigma^{2'c}_{r2jk}$ for b'^{c}_{2jk} and $1/\sigma^{2'c}_{r3jk}$ for b'^{c}_{3jk}. That means that estimates with more precision (less variance, typically associated with larger sample sizes) are given more weight in the computation of the combined (i.e., synthesized, pooled) final estimate of the relevant effect size parameter. At the second level, the effect sizes β_{2jk} and β_{3jk} from Equation 6 can be modeled as varying across participants around the study-specific mean effects, θ_{20k} and θ_{30k}:

$$\begin{aligned} \beta_{2jk} &= \theta_{20k} + u_{2jk} \text{ with } u_{2jk} \sim N\left(0, \sigma^2_{u_{2jk}}\right), \\ \beta_{3jk} &= \theta_{30k} + u_{3jk} \text{ with } u_{3jk} \sim N\left(0, \sigma^2_{u_{3jk}}\right). \end{aligned} \tag{7}$$

Figure 5 gives a graphical display of the estimated study-specific regression lines (i.e., green lines) and how the individual participants' regression lines (i.e., red lines) deviate from this for the Mason et al.'s (2014) study (note that the regression lines using the original scale are presented in Figure 5). These regression lines give an indication of the magnitude of the between-case variance in treatment effect estimates. In a next step, the effects for studies can be modeled as varying across studies:

$$\begin{aligned} \theta_{20k} &= \gamma_{200} + v_{20k} \text{ with } v_{20k} \sim N\left(0, \sigma^2_{v_{20k}}\right), \\ \theta_{30k} &= \gamma_{300} + v_{30k} \text{ with } v_{30k} \sim N\left(0, \sigma^2_{v_{30k}}\right). \end{aligned} \tag{8}$$

The meta-analyst is typically interested in the estimate of γ_{200}, referring to the average immediate treatment effect across participants and studies, and the estimate of γ_{300}, indicating the treatment effect on the time trend.

Fixed effect estimates. Using the data of the 27 SCEDs investigating the effectiveness of peer-tutoring on social outcomes scores, $\hat{\gamma}_{200}$ equals 3.74, $SE = 1.09$, $t(22.6) = 3.44$, $p = .002$, and $\hat{\gamma}_{300}$ equals 0.09, $SE = 0.19$, $t(24.4) = 0.49$, $p = .63$. Note that these are standardized, bias corrected summary estimates. In sum, these results indicate that, across all studies, peer-tutoring has a statistically significant, positive, immediate treatment effect on social outcome scores for children with learning disabilities at the .05 significance level (two-tailed testing). The treatment effect on the time trend is not statistically significant, which indicates that the trend during the treatment phase is not significantly different from the baseline trend.

Visualization of the overall average, study-specific, and case-specific regression lines applied to two participants of the Banda, Hart, and Liu-Gitz (2010) study and two participants of the Barton-Arwood (2003) study is given in Figure 6. The green line indicates the overall average regression line and is the same for the participants of the Banda et al. (2010) and Barton-Arwood (2003) studies. The blue line refers to the study-specific estimate and is the same for the participants from the same study. The red lines are participant-specific (again, note that the regression lines using the original scale are presented in Figure 6). In this way, not only the variability in treatment estimates between cases within studies is visualized but also how each individual study and each individual case deviates from the overall average treatment estimates.

Random effect estimates. In addition to the estimate of the treatment effects (i.e., fixed effects), estimates of the variability in treatment effects between cases and between studies are obtained as indicated in Equations 4 and 5 (i.e., $\sigma^2_{v_{20k}}$, referring to the between-study variance for the estimated immediate treatment effect; $\sigma^2_{v_{30k}}$, indicating the between-study variance for the estimated treatment effect on the time trend; $\sigma^2_{u_{2jk}}$, indicating the between-case variance for the estimated immediate treatment effect; and $\sigma^2_{u_{3jk}}$, referring to the between-case variance of the estimated treatment effect on the time trend). In this multilevel meta-analysis, most variability was found in the estimated immediate treatment effect between studies ($\sigma^2_{v_{20k}} = 26.79$, $SE = 9.47$, $Z = 2.83$, $p = .002$) and between cases ($\sigma^2_{u_{20k}} = 18.65$, $SE = 2.71$, $Z = 6.87$, $p < .001$). The between-study variance of the treatment effect on the time trend ($\sigma^2_{v_{30k}} = 0.50$, $SE = 0.28$, $Z = 1.79$, $p = .037$) and the between-case variance of the treatment effect on the time trend ($\sigma^2_{u_{30k}} = 2.0667$, $SE = 0.33$, $Z = 6.20$, $p < .001$) are smaller, but still statistically significant. Note that SAS 9.4 (SAS Institute Inc., SAS) provides a Z statistic and corresponding p value for statistical significance testing of the variance components estimates. Therefore, it makes the assumption that the sampling distribution of the variances is normally distributed, which

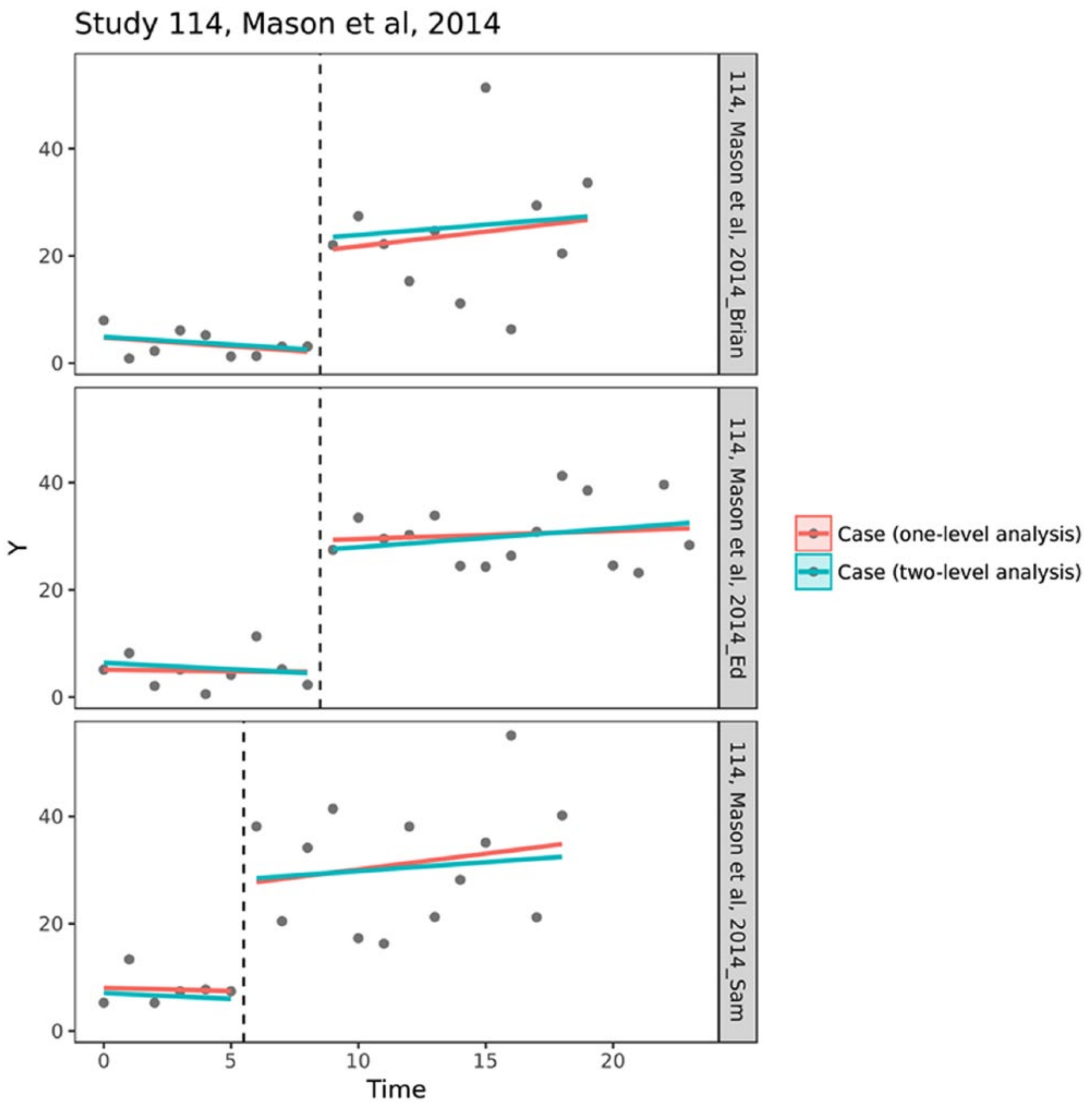

Figure 5. Graphical display of the case-specific and study-specific regression lines the three participants of the study of Mason et al. (2014).

Note. The green line indicates the study-specific regression line and the red lines are the participant-specific regression lines.

is not the case (the sampling distribution is highly positively skewed). Therefore, it is more valid to evaluate the size of the variability instead of solely relying on the Z statistic and p value. Results of a multilevel meta-analysis are typically summarized in a table as demonstrated in Table 3.

Study-level and participant-level estimates. In addition to the estimates provided in Table 3, another advantage of the multilevel meta-analytic model is that all the case-specific and study-specific effect size estimates are obtained. A large amount of variability in effect size estimates between cases and/or between studies will be reflected by a large range of case-specific and study-specific estimates. Because 27 studies with 130 cases are included in the analysis, 27 study-specific immediate treatment effects and treatment effects on time trends are estimated in addition to 130 case-specific immediate treatment effects and treatment effects on time trends. ranges from -9.73 (SE = 0.76, for Case 1 from Lorah, Gilroy, & Hineline, 2014) to 21.20 (SE = 2.33, for Case 3 from Lorah et al., 2014). This reflects the large variability between cases in the estimated treatment effect. ranges from -10.38 (SE = 0.76, Case 5, Trembath, Balandin, Togher, & Stancliffe, 2009) to 4.23 (SE = 0.79, Case 3 from Plumer, 2007). ranges from -8.21

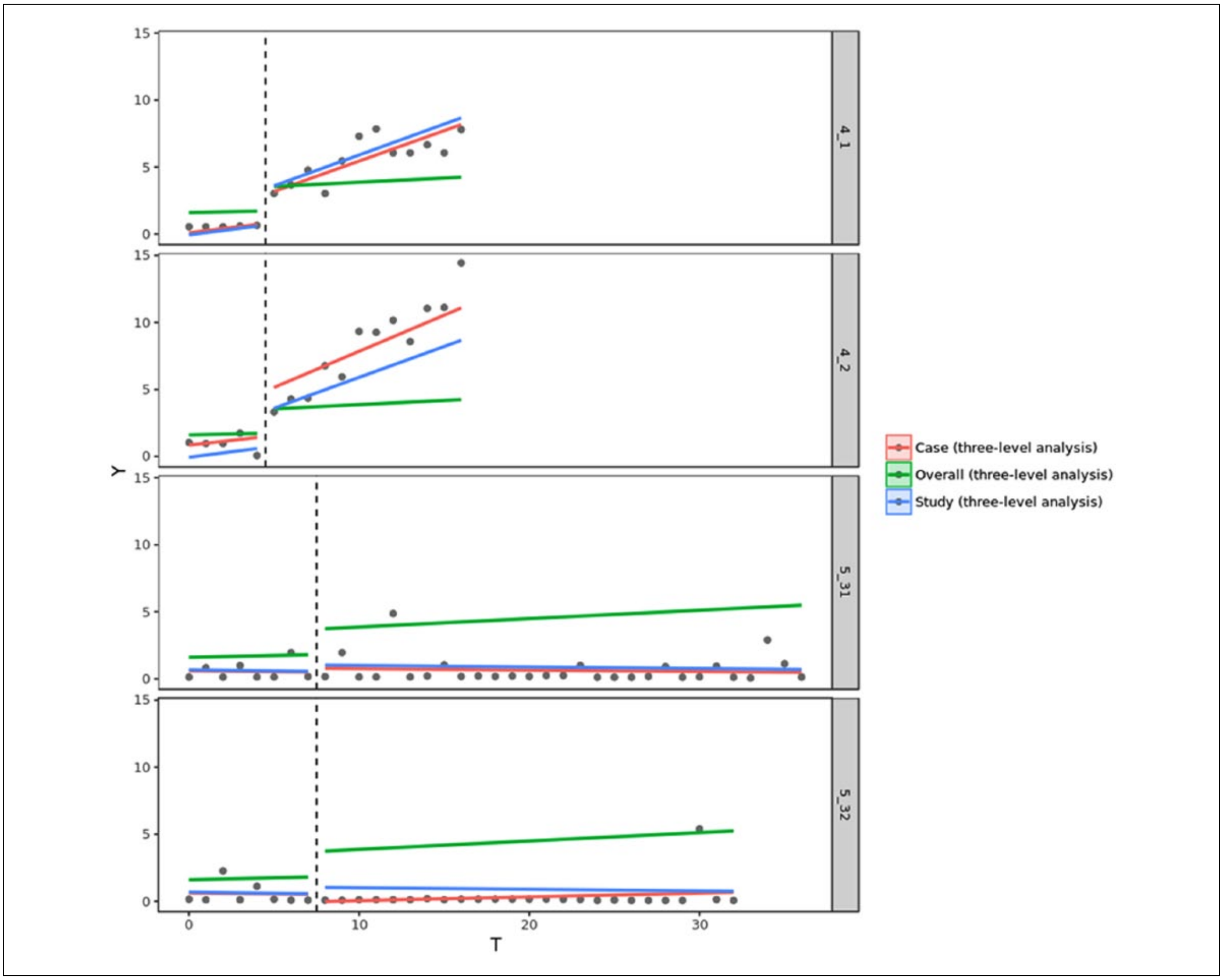

Figure 6. Graphical display of the overall average, study-specific, and case-specific regression lines applied to the two participants of the study of Banda, Hart, and Liu-Gitz (2010) and two participants of the study of Barton-Arwood (2003).
Note. The green line indicates the overall average regression line and is the same for the participants of Banda et al. (2010) and Barton-Arwood (2003). The blue line refers to the study-specific estimate and is the same for the participants from the same study. The red lines are participant-specific.

(SE = 2.24, Hibbert, Kostinas, & Luiselli, 2002) to 16.47 (SE = 2.84, Lorah et al., 2014), and has a range from -1.88 (SE = 0.51, Trembath et al., 2009) to 1.06 (SE = 0.59, Loftin, Odom, & Lantz, 2008).

Moderators. Because of the large variability (more than expected based on random error variance) in study- and case-specific effect size estimates (especially for the immediate treatment effect), it makes sense to try to explain the source of variability by adding a moderator. To illustrate this, we added age as a second-level moderator. The mean age is 8.27 years (SD = 2.90) ranging from 3 to 17 years. Age was mean-centered to avoid multicollinearity (i.e., by adding a moderator, correlation between the moderator and the other predictors might be induced). These are the modified Level 2 equations:

$$\begin{aligned} \beta_{2jk} &= \theta_{20k} + \theta_{21k} + u_{2jk} \text{ with } u_{2jk} \sim N\left(0, \sigma^2_{u_{2jk}}\right), \\ \beta_{3jk} &= \theta_{30k} + \theta_{31k} + u_{3jk} \text{ with } u_{3jk} \sim N\left(0, \sigma^2_{u_{3jk}}\right). \end{aligned} \quad (9)$$

θ_{21k} and θ_{31k} indicate the age effect on the case-specific immediate treatment effect and treatment effect on the time trend, respectively. The results can be found in Table 3. Some notable differences are that the immediate treatment effect becomes even more statistically significant, namely, γ_{200} = 3.88, SE = 1.00, t(21.6) = 3.85, p = .001. Age seems to moderate the relation between the immediate treatment effect and social outcome, γ_{210} = −0.58, SE = 0.25, t(64.1) = −2.28, p = .026. The older the participants, the less effective the peer-tutoring intervention. Age does not have a statistically significant effect on the treatment effect on the time trend. In addition, we found that age succeeds at

Table 3. Summary Multilevel Meta-Analytic Coefficients Using the Meta-Analysis of Moeyaert et al. (2018).

		Model 1			Model 2		
	Parameter	Estimate	*SE*	*p*	Estimate	*SE*	*p*
Fixed effects							
Immediate treatment effect	γ_{200}	3.74	1.09	.0023	3.88	1.00	.0009
Treatment effect on trend	γ_{300}	0.09	0.19	.6287	0.08	0.19	.6,878
Age effect on immediate treatment	γ_{210}	NA	NA	NA	−0.58	0.25	.0257
Age effect on treatment effect on time trend	γ_{310}	NA	NA	NA	0.079	0.06	.1932
Variance effects							
Study							
Immediate treatment effect	$\sigma^2_{v_2}$	26.79	9.4659	.0023	22.37	8.34	.0036
Treatment effect on trend	$\sigma^2_{v_3}$	0.50	0.2769	.0368	0.49	0.28	.0403
Case							
Immediate treatment effect	$\sigma^2_{u_2}$	18.65	2.7149	<.0001	18.65	2.73	<.0001
Treatment effect on trend	$\sigma^2_{u_3}$	2.07	0.3334	<.0001	2.062	0.33	<.0001
Residual variance	σ^2_e	1			1		

Note. Model 1 does not include the moderator variable "age" whereas Model 2 includes the moderator variable. NA = not applicable.

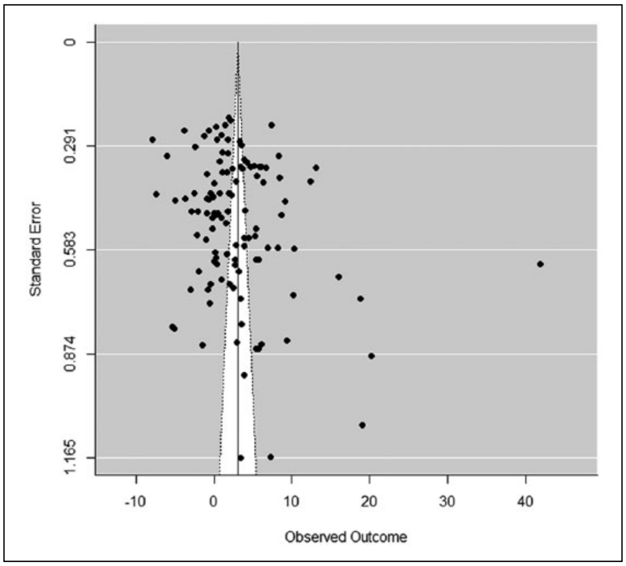

Figure 7. Funnel plot giving a graphical display of the standard error as a function of the effect size (i.e., observed outcome).

reducing the between-study variance in immediate treatment effect ($\sigma^2_{v_{20k}} = 22.37$). Table 3 indicates that a substantial amount of variability is remaining and therefore further exploration of potential moderator variables is recommended.

Publication bias. Publication bias is a common concern when performing a meta-analysis. A funnel plot can be evaluated to examine publication bias, as displayed in Figure 7. The funnel plot was created using the "metafor" package in R 3.2.5 (Viechtbauer, 2010). On the *Y*-axes the *SE*s are displayed (i.e., the root square of the inverse of the precision) and on the *X*-axes the standardized bias -corrected outcome score is displayed (i.e., the effect size). As the *SE* becomes smaller (more precision), less variability in effect size estimates is to be expected. As a consequence, ideally all the data points lie within the funnel. This general trend is not observed in current study, as there remains a substantial amount of variability in effect sizes, regardless of the precision. We can also deduce that for the larger range of values of the *SE*, there is a lack of studies, and as a consequence, we may conclude that there is some evidence for publication bias in the field of SCEDs using peer-tutoring as a treatment to increase social outcomes. For an in-depth discussion of the funnel plots, we refer readers to Sterne and Egger (2001) and Sterne and Harbord (2004).

Discussion and Extensions

The aim of this article was to introduce the basic multilevel meta-analytic model for the quantitative integration of regression-based SCED effect size estimates. The same logic can be applied to summarize other effect size estimates. An empirical demonstration together with graphical presentations and interpretations of effect size estimates was provided together with software code. The purpose was to provide applied single-case researchers and research synthesists with the necessary knowledge, conceptual understanding, and tools to independently perform the multilevel meta-analysis.

Although the focus of this article was on introducing the basic multilevel meta-analytic model, straightforward extensions can be implemented to model additional data and design characteristics. Therefore, I briefly give here an overview of the most common data and design characteristics and refer to relevant literature.

1. Continuous outcomes were assumed, but in the majority of SCED studies, the outcomes might be a sort of a count (Shadish & Sullivan, 2011). As such, a Poisson regression might be more appropriate and Poisson-based regression effect sizes can be combined (Beretvas & Chu, 2013; Declercq, Beretvas, Moeyaert, Ferron, & Van den Noortgate, 2018). A Poisson distribution makes sense as it can only take on integer values (i.e., the outcome score has values of 0, 1, 2, etc.) whereas the OLS regression outcome can have any value, integer, or fractional.
2. For demonstration purposes, linear trends in the baseline and the treatment phases were assumed. Whereas it is reasonable to assume linear (and flat) trends during the baselines, nonlinear trends might be present in the treatment phase (i.e., asymptotic trend in case there is a floor or ceiling effect, or quadratic trends). Therefore, functional forms other than linearity might be more realistic as suggested and further explored by Hembry, Bunuan, Beretvas, Ferron, and Van den Noortgate (2015).
3. Dependent errors are common in SCED data as repeated measures across time are obtained (commonly labeled as autocorrelation). Baek and Ferron (2013) discuss the issue of autocorrelation.
4. In this article, I assumed that the variance in outcome scores during the baseline phase is the same as the variance of the outcome scores during the treatment phase. The assumption of homogeneity might be violated as the data in the treatment phase might be more variable compared with baseline data. This issue of heterogeneity and methods to deal with heterogeneity are discussed by Joo, Ferron, Moeyaert, Beretvas, and Van den Noortgate (2017).
5. The studies were simplified to simple AB phase designs, but in reality, more complex SCED studies are common (e.g., alternating treatment designs and phase change reversal designs). Moeyaert, Ugille, Ferron, Beretvas, and Van den Noortgate (2014a) gave an empirical demonstration for the quantitative integration of effect sizes from different SCED types. A list of methodological work in the context of multilevel modeling if SCEDs is provided Moeyaert, Manolov and Rodabaugh (in press) for readers interested in modeling other complexities than the ones discussed in this study.

As is clear from these few examples of additional design and data characteristics, it is challenging to assume *a priori* fixed parameters (i.e., one best model) that result in the best data fit. Participants and studies are different based on their specific data and design characteristics. For instance, for some participants, a quadratic model might result in the best model fit whereas a linear model is best suited for other participants.

One suggestion for determining the best model for one's data is to first explore case-specific models. Afterward, the resulting effect sizes can be combined using multilevel meta-analysis. One promising approach is Bayesian Modeling Averaging (BMA). In the BMA framework (Leamer, 1978), we let the data speak for itself by determining which variables are most appropriate given the data. To reduce subjectivity and underestimation of model uncertainty, the decisions are automatically made for the researcher, resulting in better predictive ability. BMA involves averaging overall possible models (i.e., combination of parameters) when making inferences. BMA will result in the best set of parameters given the data per participant. This suggestion is an idea for future research.

This study presents a univariate multilevel meta-analysis as the focus is on evaluating the effectiveness of peer-tutoring interventions on one dependent variable, namely, academic outcomes. The original meta-analytic dataset of Moeyaert et al. (2018) also includes social outcomes. As it is anticipated that social and academic outcomes are correlated, a multivariate multilevel meta-analysis can be conducted. Multivariate multilevel meta-analytic models require further methodological investigation.

Author's Note

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.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the Institute of Education Sciences, U.S. Department of Education, through grant R305D150007.

References

Baek, E., & Ferron, J. M. (2013). Multilevel models for multiple-baseline data: Modeling across participant variation in autocorrelation and residual variance. *Behavior Research Methods*, *45*, 65–74. doi:10.3758/s13428-012-0231-z

Banda, D. R., Hart, S. L., & Liu-Gitz, L. (2010). Impact of training peers and children with autism on social skills during center time activities in inclusive classrooms. *Research in Autism Spectrum Disorders*, *4*, 619–625. doi:10.1016/j.rasd.2009.12.005

Barton-Arwood, S. M. (2003). *Reading instruction for elementary-age students with emotional and behavioral disorders: Academic and behavioral outcomes* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3085747)

Beretvas, S. N., & Chu, Y. (2013 April). *Handling count data outcomes trajectories in multiple-baseline design studies*. Paper presented at the annual meeting of the American Educational Research Association, Philadelphia, PA.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: John Wiley.

Card, N. A. (2012). *Applied meta-analysis for social science research*. New York, NY: Guilford.

Cools, W., Declercq, L., Beretvas, S. N., Moeyaert, M., Ferron, J. M., & Van den Noortgate, W. (2017). MultiSCED [Software]. Retrieved from http://52.14.146.253/MultiSCED/

Declercq, L., Jamshidi, L., Fernández-Castilla, B., Beretvas, S., Moeyaert, M., Ferron, J., & Van den Noortgate, W. (2018). Analysis of single-case experimental count data using the linear mixed effects model: A simulation study. *Behavior Research Methods*. Advance online publication. doi:10.3758/s13428-018-1091-y

Ferron, J. M., Bell, B. A., Hess, M. F., Rendina-Gobioff, G., & Hibbard, S. T. (2009). Making treatment effect inferences from multiple-baseline data: The utility of multilevel modeling approaches. *Behavior Research Methods*, *41*, 372–384. doi:10.3758/BRM.41.2.372

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*, 3–8. doi:10.3102/0013189X005010003

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*, 107–128. doi:10.3102/10769986006002107

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press. doi:10.1080/00220973.2014.907231

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

Hibbert, D., Kostinas, G., & Luiselli, J. K. (2002). Improving skills performance of an adult with mental retardation through peer-mediated instructional support. *Journal of Developmental and Physical Disabilities*, *14*, 119–127. doi:10.1023/A:1015263329642

Jamshidi, L., Heyvaert, M., Declercq, L., Fernández Castilla Ferron, J., Moeyaert, M., Beretvas, S. N., & Van den Noortgate, W. (2018). *A systematic review of single-case experimental design meta-analyses: Characteristics of study designs, data, and analyses*. Manuscript submitted for publication.

Joo, S., Ferron, J., Moeyaert, M., Beretvas, S., & Van den Noortgate, W. (2017). Approaches for specifying the level-1 error structure when synthesizing single-case data. *Journal of Experimental Education*. Advance online publication. doi:10.1080/00220973.2017.1409181

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, *53*, 983–997. doi:10.2307/2533558

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 https://files.eric.ed.gov/fulltext/ED510743.pdf

Leamer, E. E. (1978). *Specification searches*. New York, NY: John Wiley.

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

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

Loftin, R. L., Odom, S. L., & Lantz, J. F. (2008). Social interaction and repetitive motor behaviors. *Journal of Autism and Developmental Disorders*, *38*, 1124–1135. doi:10.1007/s10803-007-0499-5

Lorah, E. R., Gilroy, S. P., & Hineline, P. N. (2014). Acquisition of peer manding and listener responding in young children with autism. *Research in Autism Spectrum Disorders*, *8*, 61–67. doi:10.1016/j.rasd.2013.10.009

Mason, R., Kamps, D., Turcotte, A., Cox, S., Feldmiller, S., & Miller, T. (2014). Peer mediation to increase communication and interaction at recess for students with autism spectrum disorders. *Research in Autism Spectrum Disorders*, *8*, 334–344. doi:10.1016/j.rasd.2013.12.014

Moeyaert, M., Klingbeil, D., Rodabaugh, E., & Turan, M. (2018). *Multilevel meta-analysis of peer-tutoring interventions to increase academic performance and social interactions for people with special needs*. Manuscript submitted for publication.

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

Moeyaert, M., Manolov, R., & Rodabaugh, E. (in press). Meta-analysis of single-case research via multilevel models: Fundamental concepts and methodological considerations. *Behavior Modification*.

Moeyaert, M., Ugille, M., Ferron, J., Beretvas, S., & Van den Noortgate, W. (2013). The three-level synthesis of standardized single-subject experimental data: A Monte Carlo simulation study. *Multivariate Behavioral Research*, *48*, 719–748. doi:10.1080/00273171.2013.816621

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

Moeyaert, M., Ugille, M., Ferron, J., Beretvas, S., & Van den Noortgate, W. (2014b). Three-level analysis of single-case experimental data: Empirical validation. *Journal of Experimental Education*, *82*, 1–21. doi:10.1080/00220973.2012.745470

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

Parker, R. I., & Vannest, K. (2008). An improved effect size for single-case research: Nonoverlap of all pairs. *Behavior Therapy*, *40*, 357–367. doi:10.1016/j.beth.2008.10.006

Pellegrini, R. J., & Hicks, R. A. (1972). Prophecy effects and tutorial instruction for the disadvantaged child. *American Educational Research Journal*, *9*, 413–419. doi:10.3102/00028312009003413

Plumer, P. J. (2007). *Using peers as intervention agents to improve the social behaviors of elementary-aged children with attention deficit hyperactivity disorder: Effects of a peer coaching package* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3275754)

Raudenbush, S. W., & Bryk, A. S. (1985). Empirical Bayes meta-analysis. *Journal of Educational Statistics*, *10*, 75–98. doi:10.3102/10769986010002075

Rohatgi, A. (2014). *WebPlotDigitizer user manual version 3.4*. Retrieved from https://automeris.io/WebPlotDigitizer/userManual.pdf

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231–244). New York, NY: Russell Sage Foundation.

Schaalje, G. B., McBride, J. B., & Fellingham, G. W. (2001). Approximations to distributions of test statistics in complex mixed linear models using SAS Proc MIXED. In *Proceedings of the SAS Users Group International 26th annual conference* (Paper 262–26). Retrieved from http://www2.sas.com/proceedings/sugi26/p262-26.pdf

Shadish, W. R., Rindskopf, D. M., Hedges, L. V., & Sullivan, K. J. (2012). Bayesian estimates of autocorrelations in single-case designs. *Behavior Research Methods*, *45*, 813–821. doi:10.3758/s13428-012-0282-1

Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods*, *43*, 971–980. doi:10.3758/s13428-011-0111-y

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

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, *32*, 752–760. doi:10.1037/0003-066X.32.9.752

Sterne, J. A. C., & Egger, M. (2001). Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. *Journal of Clinical Epidemiology*, *54*, 1046–1055. doi:10.1016/S0895-4356(01)00377-8

Sterne, J. A. C., & Harbord, R. M. (2004). Funnel plots in meta-analysis. *The Stata Journal*, *4*, 127–141.

Sutton, A. J., Abrams, K. R., Jones, D. R., Sheldon, T. A., & Song, F. (2000). *Methods for meta-analysis in medical research*. Chichester, UK: John Wiley.

Talbott, E., Maggin, D. M., Van Acker, E. Y., & Kumm, S. (2018). Quality indicators for reviews of research in special education. *Exceptionality*, *26*, 245–265 doi:10.1080/09362835.2017.1283625

Trembath, D., Balandin, S., Togher, L., & Stancliffe, R. J. (2009). Peer-mediated teaching and augmentative and alternative communication for preschool-aged children with autism. *Journal of Intellectual & Developmental Disability*, *34*, 173–186. doi:10.1080/13668250902845210

Ugille, M., Moeyaert, M., Beretvas, S. N., Ferron, J., & Noortgate, W. (2012). Multilevel meta-analysis of single-subject experimental designs: A simulation study. *Behavior Research Methods*, *44*, 1244–1254. doi:10.3758/s13428-012-0213-1

Ugille, M., Moeyaert, M., Beretvas, S. N., Ferron, J., & Van Den Noortgate, W. (2013). Bias corrections for standardized effect size estimates used with single-subject experimental designs. *Journal of Experimental Education*, *45*, 547–559. doi:10.1080/00220973.2013.813366

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

Van den Noortgate, W., & Onghena, P. (2003b). Hierarchical linear models for the quantitative integration of effect sizes in single-case research. *Behavior Research Methods, Instruments, & Computers*, *35*, 1–10. doi:10.3758/BF03195492

Van den Noortgate, W., & Onghena, P. (2007). The aggregation of single-case results using hierarchical linear models. *The Behavior Analyst Today*, *8*, 196–209. doi:10.1037/h0100613

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. doi:10.1080/17489530802505362

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. Retrieved from http://www.jstatsoft.org/v36/i03/