Behavior Research Methods (2020) 52:2460–2479
https://doi.org/10.3758/s13428-020-01417-0

Assessing consistency of effects when applying multilevel models to single-case data

Rumen Manolov[1] · John M. Ferron[2]

Published online: 21 May 2020

Abstract

In the context of single-case experimental designs, replication is crucial. On the one hand, the replication of the basic effect within a study is necessary for demonstrating experimental control. On the other hand, replication across studies is required for establishing the generality of the intervention effect. Moreover, the "replicability crisis" presents a more general context further emphasizing the need for assessing consistency in replications. In the current text, we focus on replication of effects within a study, and we specifically discuss the consistency of effects. Our proposal for assessing the consistency of effects refers to one of the promising data analytical techniques, multilevel models, also known as hierarchical linear models or mixed effects models. One option is to check, for each case in a multiple-baseline design, whether the confidence interval for the individual treatment effect excludes zero. This is relevant for assessing whether the effect is replicated as being non-null. However, we consider that it is more relevant and informative to assess, for each case, whether the confidence interval for the random effects includes zero (i.e., whether the fixed effect estimate is a plausible value for each individual effect). This is relevant for assessing whether the effect is consistent in size, with the additional requirement that the fixed effect itself is different from zero. The proposal for assessing consistency is illustrated with real data and is implemented in free user-friendly software.

Keywords Single-case design · Replication · Consistency · Multilevel models · Random effects

Introduction

Single-case experimental designs (SCEDs) are research designs that involve the study of one or several individuals longitudinally, with multiple measurements taken under different conditions manipulated by the researcher. SCEDs offer the possibility to carry out methodologically rigorous studies for gathering evidence on the effect of interventions (Barlow et al., 2009). SCEDs have been recognized as useful in a variety of contexts including special education (Ledford & Gast, 2018), neuropsychological rehabilitation (Tate & Perdices, 2019), sport psychology (Barker et al., 2011), and biomedicine (Janosky et al., 2009). The field has witnessed developments in terms of assessing methodological quality (Ganz & Ayres, 2018), data analysis (Kratochwill & Levin, 2014), and meta-analysis (Maggin et al., 2017), as well as reporting (Tate et al., 2016). Nevertheless, several challenges remain, such as choosing among many data analytical options (Manolov & Moeyaert, 2017) and discussing the importance of randomization (Kratochwill & Levin, 2010; Ledford, 2018) and replication (Lanovaz et al., 2019).

The aim of the current study is to propose a way of assessing consistency in data features and consistency of effects when performing a multilevel analysis of single-case data. Given that assessment consistency is based on the need for replication in single-case research, we first discuss the concepts of replication and consistency, highlighting their relevance and recent salience. We then provide a rationale for focusing on multilevel models as a data analytical technique.

Replication and consistency

A recent special issue of *Perspectives on Behavior Science* (Hantula, 2019) focused on the "replicability crisis" in psychology, and how behavior analysts can thoughtfully proceed

✉ Rumen Manolov
rrumenov13@ub.edu

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

[2] Department of Educational and Psychological Studies, University of South Florida, Tampa, FL, USA

in their use of SCEDs. In the SCED context, within-study replication is relevant for internal validity, although it is only one of several aspects to consider (Ganz & Ayres, 2018; Perdices et al., 2019; Wendt & Miller, 2012). Specifically, the iterative manipulation of the independent variable and the subsequent changes observed in the dependent variable increase the confidence that these changes are not due to external factors (Horner et al., 2005) such as history and maturation (Petursdottir & Carr, 2018). In order to document experimental control, the covariation between changes in the behavioral pattern and the introduction (and withdrawal) of the intervention is to be observed at least three times, with more specific recommendations available according to the specific SCED used (What Works Clearinghouse, 2020).

Two kinds of replication can be distinguished with a SCED study. "Direct replication", or "within-subject replication", takes place in a reversal/withdrawal, multiple-baseline design, or an alternating treatments design (Horner et al., 2005; Tincani & Travers, 2019). Additionally, "systematic replication", or "inter-subject replication", can be achieved within a study (e.g., replication of a reversal/withdrawal or an alternating treatments design across participants, replication across settings of a multiple-baseline design across participants) or across studies (Horner et al., 2005; Kennedy, 2005).

When dealing with direct replication, one of the relevant concepts is consistency (Lane et al., 2017; Ledford, 2018). Although consistency has been highlighted especially in the context of visual analysis (What Works Clearinghouse, 2020), there have also been recent proposals for its quantification (Tanious, De, Michiels, et al., 2019b; Tanious, Manolov, et al., 2019c). Specifically, two types of consistency can be distinguished, both visually and quantitatively: consistency of measurements from similar phases, and consistency of effects (e.g., when comparing data points from adjacent phases).

We consider that it is necessary to distinguish between a successful replication of an effect and a successful and consistent replication. Whether a "basic effect" (Horner & Odom, 2014) is present is an assessment that is usually performed visually, dealing with several data features, such as level, trend, variability, overlap, and immediacy (Ledford et al., 2019; Maggin et al., 2018; What Works Clearinghouse, 2020). Subsequently, several attempts to replicate the basic effect take place, and an evaluation is performed regarding whether the replication was successful (i.e., whether a functional relation or experimental control is documented). However, suppose that we proceed quantitatively, and the focus of the quantification is put on the immediate effect because trends are not expected: the difference between the mean of the last three baseline data points and the first three intervention phase data points could be computed (Horner & Kratochwill, 2012; Michiels & Onghena, 2019a). On the one hand, if the immediate effects for each participant in a study are all greater than zero (or than a minimally relevant effect), this would be indicative of a successful replication, assuming there are no other data features (e.g., trend, variability) that suggest the contrary. On the other hand, if the values of the immediate effect are similar (e.g., there are small deviations from the average effect, which is greater than zero or than a minimally relevant effect), this would be indicative of a successful replication with a consistent immediate effect. In the following text, we focus on multilevel models, and we first discuss a definition for successful replication before presenting our main proposal for a definition of a successful and consistent replication.

Focus on multilevel modeling

Multilevel models are one of the promising analytical alternatives for SCED data analysis (Van den Noortgate & Onghena, 2007), and they have been recommended in domains such as education (Dedrick et al., 2009), experimental psychology (DeHart & Kaplan, 2019), and aphasiology (Wiley & Rapp, 2019). Multilevel models were chosen as the focus of the current text, as they are applicable to different SCEDs and enable one to take multiple data features into account (Pustejovsky et al., 2014; Shadish et al., 2013). For instance, unlike nonoverlap measures, multilevel models take autocorrelation into account (Baek & Ferron, 2013). And in contrast to between-case standardized mean difference (Shadish et al., 2014) and the log response ratio (Pustejovsky, 2018), they do not assume absence of trend or require detrending. Moreover, multilevel models do not preclude the use of visual analysis (Davis et al., 2013).

The focus of the current text is on the evidence obtained in a single study, using a SCED. This initial clarification is important for two reasons. One the one hand, replication in the SCED context can refer both to repeated demonstrations of a basic effect (e.g., a difference between two adjacent phases) in the same study (Ninci, 2019) and to the replication of effects across studies in relation to the way in which a practice can be established as being "evidence-based" (Jenson et al., 2007; Schlosser, 2009). On the other hand, multilevel models, which are the focus of the current text, have a noteworthy application for meta-analysis (Moeyaert, 2019; Van den Noortgate & Onghena, 2003a, 2003b). In the current text, we focus here on within-study replication and the use of multilevel models as in studies using multiple-baseline designs (Ferron et al., 2009). At the within-study level, the multilevel model usually includes two levels, whereas at the across-studies levels, it usually includes at least three levels (Moeyaert et al., 2014a), although several variations are possible.

In the next section we discuss several possible ways in which consistency or results could be assessed when using multilevel models. We then offer a proposal and illustrate it with real data.

Defining successful replication in the context of a multilevel model

A ratio of effects to no effects

The "replicability crisis" has been linked to the misuse and abuse of null hypothesis testing (Branch, 2019) and to the fact that *p*-values do not inform about the likelihood to replicate the effect observed in a given sample (Killeen, 2019). As stated previously, in the SCED context, the presence or absence of a basic effect is usually determined by visual analysis rather than by means of statistical tests (Maggin et al., 2018), and this effect has to be replicated several times within the same study (What Works Clearinghouse, 2020). For the most commonly used designs – multiple-baseline and reversal/withdrawal (Shadish & Sullivan, 2011) – the requirement is for three replications. However, the recommendation of three demonstrations of a basic effect (for direct replications), just like the requirement for the amount of evidence required for calling a practice "evidence-based" (see the 5-3-20 rule in Horner & Kratochwill, 2012), more closely related to systematic replications, do not take into account the number of attempts for replication that did not yield the expected positive result. Following Kratochwill et al. (2018), it is possible to distinguish between a "negative result" (absence of demonstration of an effect or lack of evidence for effectiveness) and a "negative effect" (an iatrogenic effect of the intervention). The implications of these two different kinds of unexpected and undesired results are not identical. While a negative effect may more clearly provide evidence against an intervention, a lack of a positive result may lead to introducing methodological modifications (Tincani & Travers, 2018) or to identified relevant moderator variables related to the characteristics of participant and/or the target behavior (Ledford et al., 2016). Such considerations are only possible if selective reporting of positive results does not take place (Shadish et al., 2016; Simmons et al., 2011).

In summary, a given practice can be labeled as evidence-based, potentially evidence-based, neutral/mixed effects, insufficient evidence, or negative effects according to the number of methodologically rigorous studies and their results (Cook et al., 2015). Specifically, for direct replication in the SCED context, it has been suggested that a ratio of at least 3:1 effects to no effects (with no evidence for negative effects) is necessary for demonstrating experimental control (Cook et al., 2015; Maggin et al., 2013). Incidentally, the 3:1 ratio suggested resembles the historically used critical ratio of three (Garrett, 1937), which usually related a mean difference to its standard error (e.g., Nolte, 1937). Before following the 3:1 ratio, it is necessary to define what an "effect" is; the following paragraphs in this section deal with this aspect.

Defining an "effect"

It may not be straightforward to define what an effect is when performing a visual analysis (see Wolfe et al., 2019), but we will not discuss this here, given that the focus is on multilevel models. In terms of quantifying, it may be more straightforward to objectively define an "effect", but it is still not a flawless process. At the outset, we discard grounding the definition of an "effect" on the estimate of the fixed effect (e.g., whether it is greater than zero), because it only refers to the average and not to each of the replications. Moreover, we also discard using statistical significance as the sole basis for defining an effect. Apart from the usually mentioned interpretative drawbacks of a *p* value (Gigerenzer, 2004; Nickerson, 2000), it is not clear that any extrapolation to a population is reasonable in absence of random sampling of individuals (Edgington & Onghena, 2007).

An initial option is to put the focus on the sign of the empirical Bayes estimates obtained for the individual treatment effects (Ferron et al., 2010). An individual treatment effect of the correct sign (indicating an improvement) would be interpreted as an "effect". Subsequently, if the ratio of individual effects with the predicted sign, to the effects with the opposite sign, is at least 3:1, this could be interpreted as sufficient evidence for direct replication. Additionally, borrowing the logic of the difference between p_{rep} (replication of the correct sign) and $p_{support}$ (replication of an effect size of a certain size or more; see Sanabria & Killeen, 2007), a minimally relevant difference can be determined prior to gathering the data for labeling the effect as significant. However, we consider that the focus on the point estimate of the individual treatment effect may not be justified, given that these estimates are biased (Ferron et al., 2010).

In order to take into account the precision of the estimates, a more stringent and probably more defensible option would be to count as an "effect" the individual treatment effects whose confidence intervals are entirely on the predicted side of 0. That is, only intervals not containing zero would be considered positive effects. Analogously, the confidence interval could be required to exceed a prespecified minimally important difference entirely. Therefore, the definition of a successful replication would be to require a 3:1 ratio of confidence intervals of the individual treatment effects not including zero or a minimally important difference.

Obtaining individual treatment effects in multilevel models

When using multilevel models, it is necessary to construct a design matrix that represents the kind of effect that the researcher is interested in modeling (Moeyaert et al., 2014b). In order to obtain the individual treatment effect estimates and their confidence intervals, the dummy variable

representing the phase has to be included as a random effect but not as a fixed effect (Ferron et al., 2010; Van den Noortgate and Onghena, 2003a). The confidence intervals are constructed assuming normal distributions and equal within-phase variances and covariances, on the basis of a non-central *t*-distribution (Van den Noortgate & Onghena, 2003a). Ferron et al. (2010) recommend constructing the confidence intervals on the basis of the Kenward-Roger method for estimating the degrees of freedom.

It is worth noting that, even if all individual treatment effects are greater than zero (or than the minimally relevant difference), this does not mean that they are similar in value. Thus, following this option, we would have evidence on whether the replication is successful, but not whether it is consistent. We deal with consistency of effects in the following section.

Defining successful and consistent replication in the context of a multilevel model

More loosely related antecedents: A review of quantifications of heterogeneity

The current text deals mainly with one of the two types of consistency: consistency of effects. In order to obtain some overall indication of the difference between conditions and to gain statistical power, "internal meta-analysis" of the results obtained in a single study has been suggested (Goh et al., 2016; Hales et al., 2019). In relation to meta-analysis, it could be considered that it provides a way to measure consistency or heterogeneity of effects (Swan et al., 2020). Specifically, a possible quantification of the degree of (lack of) consistency could stem from the heterogeneity test and quantifications. However, the Q-test can be expected to have low statistical power when few effect sizes (here, direct replications) are quantitatively integrated (Lipsey & Wilson, 2001). Additionally, a drawback related to the descriptive quantification known as I^2 (the proportion of true variance in effect sizes with respect to the total observed variance) is that it is only a relative measure and may not be sufficiently informative (Borenstein et al., 2017). Therefore, it seems that these two options cannot be meaningfully borrowed from the general context of meta-analysis and adopted for quantification of consistency of effects at the within-study level.

Two of the analytical procedures proposed for SCED data are worth noting, because they (a) are directly applicable to studies including several participants, and (b) incorporate quantifications that can be useful for assessing consistency of effects as an indicator of the degree to which direct replication has been achieved. The between-case standardized mean difference (BC-SMD; Hedges et al., 2012, 2013) yields, among other quantifications, an "intraclass correlation" (ICC), interpreted as the amount of variability across participants as a proportion of the whole variability (within and across participants). Therefore, this value could be understood to quantify the degree to which the data patterns are not consistent, with 0.3 as a possible cutoff value indicating consistency (Hedges et al., 2012). The ICC in the BC-SMD context can be understood as representing both the consistency of data in similar phases and the consistency of effects, because even if the average difference were the same for all participants, the ICC would not be equal to zero unless the phase means were also the same across participants. Thus, it is not a pure quantification of consistency of effects.

In the context of multilevel models, an ICC can also be computed, with a similar interpretation as for the BC-SMD (see Dixon & Cunningham, 2006, for several interpretations). Actually, the ICC is usually computed for a null (also called unconditional or intercept-only) model without predictors, in order to verify whether a multilevel model is needed, i.e., whether there are relevant dependencies to be modeled (Gage & Lewis, 2014). Thus, its use, after the definitive model with predictors is built, is not that common.

More closely related antecedents: Quantifications of consistency

Tanious, De, Michiels, et al. (2019b) propose a quantification of consistency of effects, called CONEFF, referring to five data aspects, as present in the What Works Clearinghouse (2020) Standards: change in level (standardized mean difference), change in trend (using ordinary least squares estimation), change in variability (variance ratio), immediacy of the effect (the last three baseline phase measurements compared with the first three intervention phase measurements), and overlap between data from adjacent phases (using the Nonoverlap of All Pairs; Parker & Vannest, 2009). Actually, CONEFF could be applied to other ways of quantifying these five data features. Here we focus on the assessment of consistency of the change in level and change in slope, in the context of a multilevel model. As a strength of the current proposal, the use of multilevel models eliminates the ambiguity regarding exactly how to operatively define data features such as overlap and trend, both with multiple definitions suggested in the SCED context (see Parker et al., 2011, and Manolov, 2018, respectively).

A quantification of consistency of data in similar phases, called CONDAP, has been suggested for several SCEDs (Tanious, De, Michiels, et al., 2019b; Tanious, Manolov, et al., 2019c). CONDAP can be accompanied by a randomization test in case randomization is present in the design (Tanious, De, & Onghena, 2019a). CONDAP is based directly on the data, without referring to any analytical procedure or representation such as a mean line or a trend line. In contrast, we propose an assessment of the consistency of data in similar

phases related to the estimates of the intercept and baseline trend, according to a multilevel model. The aim is to fully benefit from the output of a multilevel analysis (e.g., interpreting individual treatment effects and random effects). Nevertheless, if desired, an additional quantification such as CONDAP can be used for an assessment of consistency of data patterns in similar phases that is not based on modeling.

Alternatives for quantifying consistency in the context of a multilevel model

Discussing initial options

In the context of multilevel models, when the immediate change in level and the change in slope are modeled as random effects, it is possible to compute the variance in these effects. These variance estimates could then be used as an indicator of lack of consistency. One approach for doing so would be to argue that if a variance is not statistically significant, then a random effect is not necessary in the model, because there is not sufficient variability across participants in the treatment effect. However, there are three reasons why we do not recommend using the statistical significance of the variance as a criterion. First, there are different ways to assess the importance of a random effect statistically: via a Z test under the assumption that the sampling distribution of the variances is normal (Moeyaert, 2019), or comparing the deviance values (−2 times the log likelihood) of the models with and without the random effect via a chi-square test (Hox, 2010). These two tests need not necessarily coincide, and both are suspect with small sample sizes, because the variance estimates are biased in such contexts (Ferron et al., 2009). Second, not rejecting the null hypothesis does not justify drawing a conclusion about similarity (Gigerenzer, 2004), and it is not the same as performing a test of statistical equivalence (Tryon, 2001). Third, a summary measure such as the variance and the evaluation of statistical significance seem too general for assessing consistency across individuals, as they collapse all the information about the variation in a single value (the estimate or the p value). In contrast, in the SCED context, it is recommended that the information be summarized in such a way as to maintain the information about each individual (Hagopian, 2020), which is also well aligned with some statistical approaches for contrasting hypothesis for all participants, rather than on average (Klaassen, 2020). Accordingly, the proposal that we make in the following section allows us to represent how much each individual effect differs from the average, rather than how much all individuals, on average, differ from the average.

To aid in interpreting the variance, and so as not to focus exclusively on its associated p value, a coefficient of variation could be computed for each of the effects: immediate change in level and change in slope. The numerator would be the square root of the estimated variance of the effect, and the denominator would be the absolute value of the corresponding fixed effect (i.e., estimated average). The coefficient of variation can be expressed as a percentage, but unlike the ICC or I^2, it is relative to the average effect estimated, which may lead to more meaningful interpretations regarding whether this variability is considerable. In order to reference the coefficient of variation as a quantification of how consistent (or actually, not consistent) the effect is, the fixed effect estimate should be indicative of an effect being present.

As a limitation of the use of the coefficient of variation, it must be mentioned that a specific and universal cutoff point for a "small" coefficient of variation (and sufficient consistency to be interpreted as a successful and consistent replication) does not exist. A second, and more important, limitation is that there is evidence that the variance estimates can be biased[1] for fewer than five participants in the study (Ferron et al., 2009; Moeyaert et al., 2017). A third limitation is that, if the fixed effect estimate is very small (e.g., close to zero), a large coefficient of variation can be expected, and this would reduce its informative value. Therefore, the coefficient of variation needs to be interpreted with caution. As an alternative, we now present our main proposal.

A proposal for assessing consistency of individual effects

An alternative to using the variance estimate as the basis for assessing consistency would be to use the random effect estimates. This proposal is similar to the previously mentioned possibility for assessing replication in that it is based on confidence intervals. For assessing replication, we focused on the confidence intervals for the individual intervention effects. In contrast, here we focus on the confidence intervals for the random effects (i.e., the difference between the fixed effect estimate and the individual treatment effect estimate). Specifically, it is possible to check how many of the confidence intervals for the random effects include 0. In this case, a value of zero for the random effect would represent an individual treatment effect equal to the fixed effect estimate (i.e., the average for all participants). In that sense, if the confidence interval

[1] Ferron et al. (2009) used a restricted maximum likelihood estimation applied to data including an immediate and sustained change in level, and report that the between-participants variance in the treatment effect was overestimated. In contrast, Moeyaert et al. (2017) generated data including both an immediate change in level and a change in trend, and report that the between-participants variance in the *immediate* treatment effect was underestimated, for both full and restricted maximum likelihood estimation. For the evaluation of consistency, underestimating the variance of the effect would induce a false "evidence" for consistency (i.e., a false positive), whereas overestimating the variance would induce false "evidence" against consistency (i.e., a false negative). The former is likely to be considered more detrimental, considering the alpha and beta error rates that are usually considered acceptable (Cohen, 1992).

Table 1. Results from applying multilevel models representing change in level, using the lme4 package

Aspect	Average estimate	Standard error	Standard deviation
Model. L1 for the Lambert et al. (2006) data			
Baseline level	7.00	0.34	0.76
Change in level	−5.79	0.43	0.62
Model.L2 for the Lambert et al. (2006) data			
Baseline level	6.56	0.63	1.74
Change in level	−4.85	0.54	1.53
Model.S1 for the Sherer and Schreibman (2005) data			
Baseline level	15.64	9.76	23.53
Change in level	21.91	9.71	23.06

Note. The average estimate represents the fixed effect, whereas the standard deviation represents the random effect

for a random effect includes 0, then it would be plausible for the individual treatment effect to be equal to the average. This method of assessing consistency is strengthened by having longer observation series with less error variance, because studies that are designed in this manner will tend to have more precise estimates of the random effects (i.e., narrower confidence intervals that all include 0 make a stronger argument for consistency). With this option, it would still be necessary to check that the fixed effect estimate exceeds zero or a minimally relevant value. Note that for obtaining the estimates of the random effects for the treatment effect, it is necessary to include the dummy variable representing the phase both in the fixed and in the random part of the equation.

Once the number of positive and consistent effects is tallied, two quantifications are possible. On the one hand, it can be checked whether the ratio of effects to no effects meets or exceeds 3:1. On the other hand, the percentage of positive and consistent effects can be computed. Obviously, the 3:1 ratio corresponds to 75% of the confidence intervals for the random effects including 0.

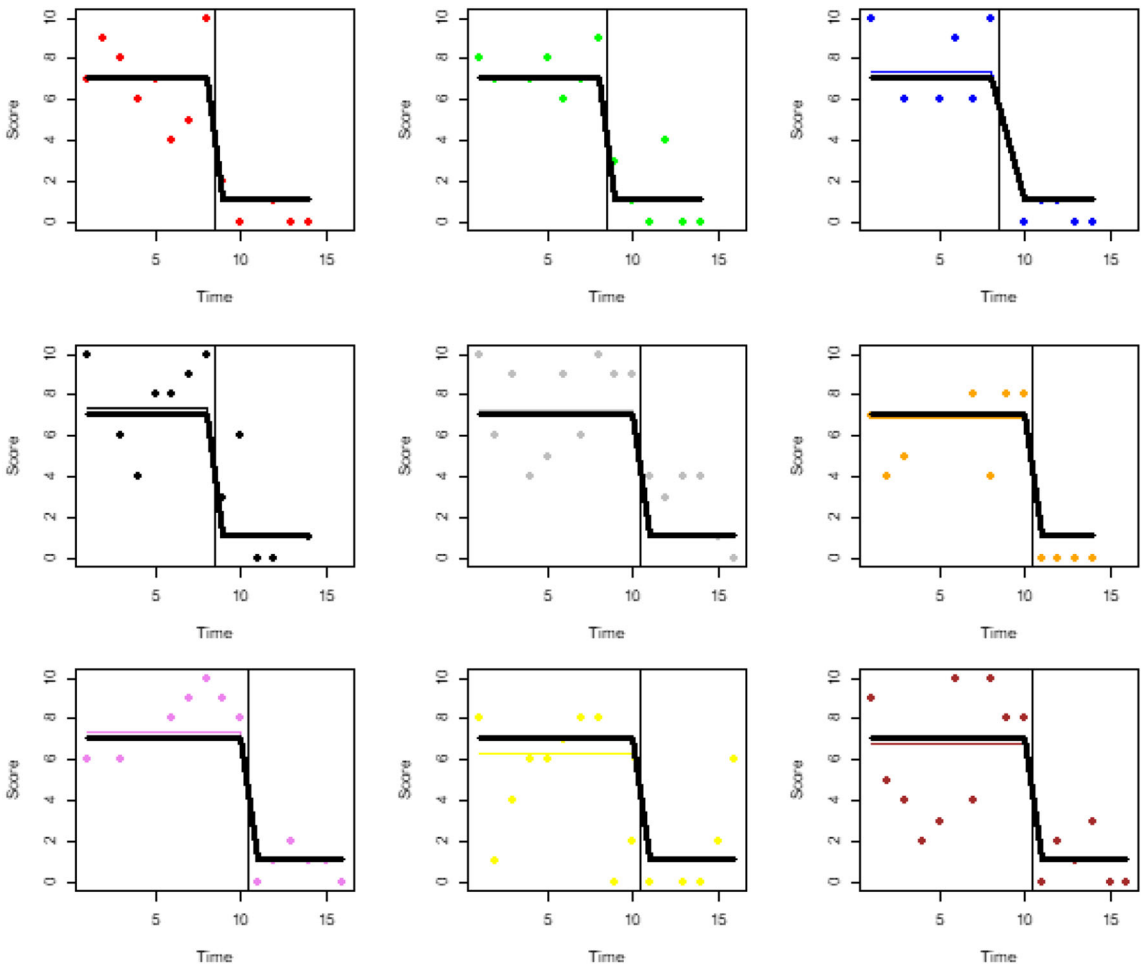

Fig. 1 Graphical representation of the multilevel model representing change in level, applied to the A_1-B_1 comparisons from the Lambert et al. (2006) data

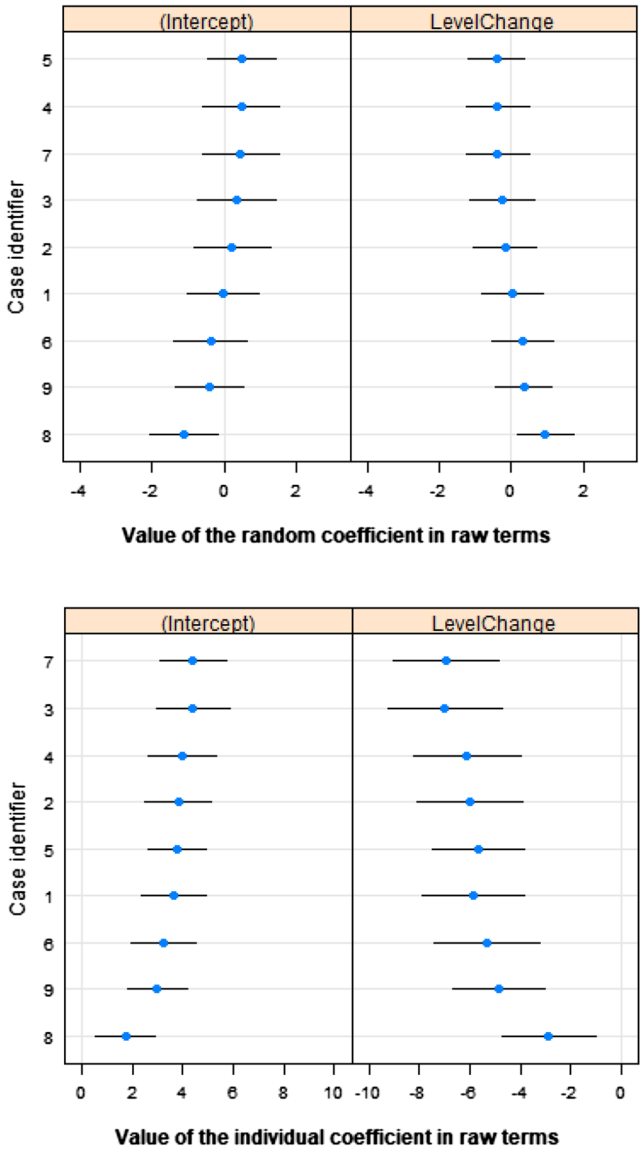

Fig. 2 The upper panel includes the caterpillar plot for the random effects and their confidence intervals as obtained via a multilevel model and including only change in level, for the A_1-B_1 comparison from Lambert et al. (2006). The lower panel includes the empirical Bayes estimates of the individual treatment effects

Illustrating the assessment of consistency of effects: Lambert et al. (2006) data

The data on disruptive behaviors, gathered by Lambert et al. (2006) following an ABAB design replicated over nine participants, has been used in several articles that present and compare different analytical options (e.g., Michiels & Onghena, 2019b ; Moeyaert et al., 2014a; Peng & Chen, 2015; Shadish et al., 2014). In terms of a quantification of consistency, we will discuss the results of two of these articles. Shadish et al. (2014) applied the BC-SMD and obtained a bias-adjusted standardized mean difference equal to −2.51 and, more importantly for the current aim, an ICC equal to .03, suggesting the almost all the variability in scores is within participants, indicating consistent results across participants. Moeyaert et al. (2014a) applied several multilevel models, and here we focus on the model that quantifies the average difference in level, without considering trend or autocorrelation, presenting quantifications separately for the A_1-B_1 comparison and for the A_2-B_2 comparison (Model 1B in Moeyaert et al., 2014a). For the change in level in the A_1-B_1 comparison, the variance reported is equal to 0, suggesting a strong consistency in the effect. For the change in level in the A_2-B_2 comparison, the variance reported is equal to 1.02, with an associated p value of .148, indicative of lower consistency as compared with the effect in the A_1-B_1 comparison. The software used by Moeyaert et al. (2014a) for obtaining the estimates is SAS 9.3. In order to be able to use a caterpillar plot to represent the random effects, we used the R package *lme4* (https://cran.r-project.org/web/packages/lme4/index.html). The data file used for the illustration provided here can be downloaded from https://osf.io/p3bna/, where there is also a time series line plot representing the measurements obtained by Lambert et al. (2006).

For this initial illustration ("Model.L1"), we apply a multilevel model which includes only a dummy variable representing phase and treats this dummy variable as a random effect. In that sense, the estimates obtained are the average baseline level and the average change in level when the intervention is introduced (as fixed effects) and the between-case variance of these effects. The numerical results can be found in Table 1. Additionally, the individual empirical Bayes estimates for level and change in level were obtained, ranging from 5.86 to 7.47 for the baseline level and from −6.20 to −4.86 for the change in level. The graphical representation of the model is shown in in Fig. 1.

For the A_1-B_1 comparison, we obtained the caterpillar plot of the random effects represented in Fig. 2, upper panel. It can be seen that eight out of nine confidence intervals (88.89% or a ratio of 8:1) include the fixed effect estimate (equal to −5.79 using the *lme4* package vs. −5.66 reported by Moeyaert et al., 2014a, using SAS). Additionally, the lower panel of Fig. 2, including the empirical Bayes estimates of the individual treatment effects (LevelChange), indicates that all nine point estimates suggest a reduction. Actually, eight of the individual effects exceed a reduction of five disruptive behaviors. However, a cutoff value for a minimally relevant difference would ideally be established prior to gathering the data. The coefficient of variation, dividing the square root of the variance by the estimate of the fixed effect would be $100 \times (0.62423/|-5.7955|)=10.77\%$.

For the A_2-B_2 comparison ("Model.L2"), the graphical representation is shown in Fig. 3, whereas the numerical results regarding the fixed and random effects can be found in Table 1. Additionally, the individual empirical Bayes estimates for level and change in level were obtained, ranging from 4.18 to 8.52 for the baseline level and from −6.45 to −3.37 for the change in level.

Regarding the assessment of consistency, the caterpillar plot of the random effects is represented in Fig. 4, upper panel. It can be seen that five out of nine confidence intervals

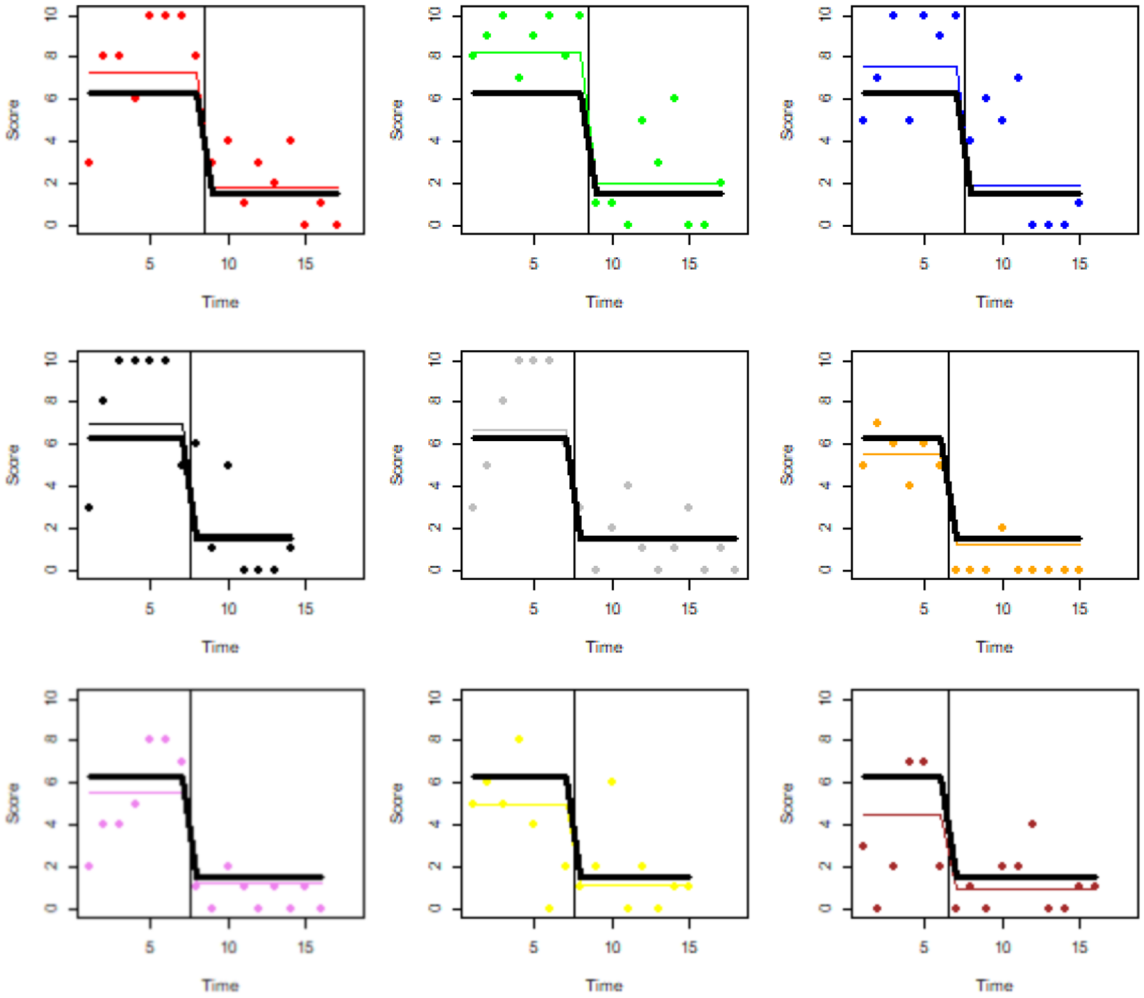

Fig. 3 Graphical representation of the multilevel model representing change in level, applied to the A_2-B_2 comparisons from the Lambert et al. (2006) data

(55.56% or a ratio of 1.25:1) include the fixed effect estimate (equal to −5.06 using the *lme4* package vs. −5.08 reported by Moeyaert et al., 2014a, using SAS). Additionally, the lower panel of Fig. 4, including the empirical Bayes estimates of the individual treatment effects, indicates that all nine point estimates suggest a reduction. However, greater variability in the A_2-B_2 effects is visible as compared with the A_1-B_1 effects in the lower panel of Fig. 2. Accordingly, the coefficient of variation, dividing the square root of the random effect by the estimate of the fixed effect, is larger than for the A_1-B_1 comparison: $100 \times (1.2349/|-5.061032|) = 24.40\%$.

Illustrating the assessment of consistency of effects: Sherer and Schreibman (2005) data

In order to illustrate the results for a data set with lower consistency, we use the data on appropriate speech gathered by Sherer and Schreibman (2005) using a multiple-baseline design across participants, and included in the illustration of multilevel modeling for meta-analysis by Moeyaert et al. (2014a). Just as for the previous illustration, we apply a multilevel model ("Model.S1") which includes only a dummy variable representing phase and treats this dummy variable as a random effect. The fixed and random effects can be found in Table 1. Additionally, the individual empirical Bayes estimates for level and change in level were obtained, ranging from −0.47 to 49.83 for the baseline level and from 1.10 to 56.82 for the change in level. The graphical representation of the model is shown in in Fig. 5.

Figure 6, upper panel, includes the caterpillar plot, according to which only one of the six confidence intervals (16.67%) includes the fixed effect estimate. The lower panel illustrates the variability in the individual treatment effects, with two of them very close to zero. Similarly, according to the coefficient of variation, dividing the square root of the random effect by the estimate of the fixed effect, there is considerable variation: $100 \times (23.059/|21.91|) = 105.24\%$. This variability is related to the presence of two different profiles of participants in the Sherer and Schreibman (2005) study: responders and nonresponders. Here we used the data

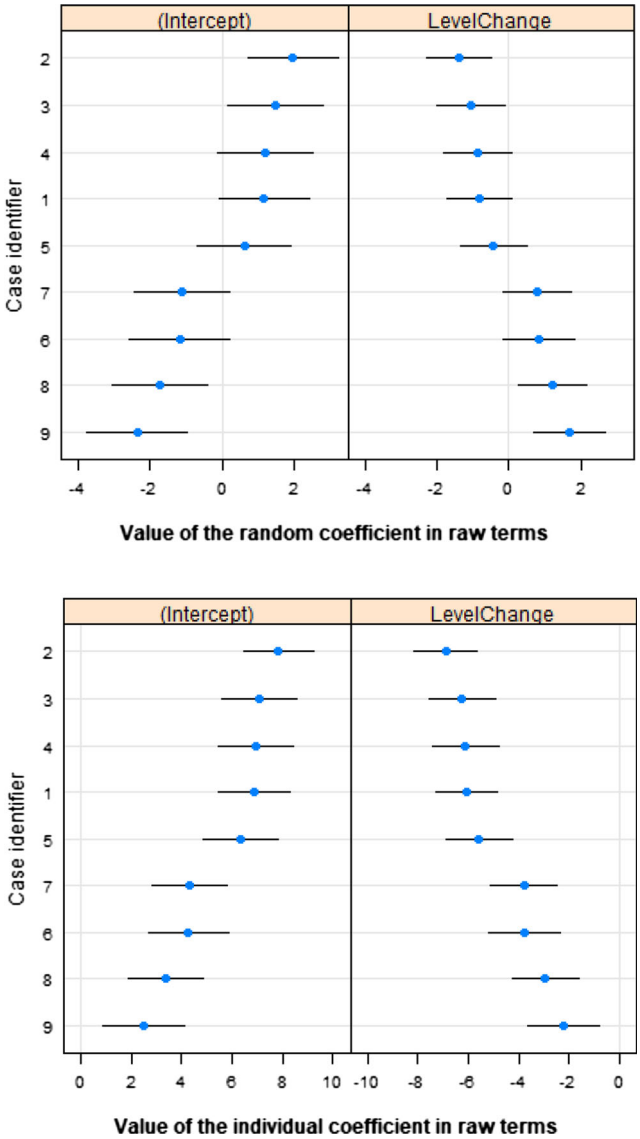

Fig. 4 The upper panel includes the caterpillar plot for the random effects and their confidence intervals, as obtained via a multilevel model including only change in level, for the A_2-B_2 comparison from Lambert et al. (2006). The lower panel includes the empirical Bayes estimates of the individual treatment effects

in order to illustrate a study with lack of consistency, without suggesting that it is necessarily meaningful to integrate the results of all participants quantitatively. The data file used for the illustration provided here can be downloaded from https://osf.io/p3bna/, where there is also a time series line plot representing the measurements obtained by Sherer and Schreibman (2005).

Additional illustration with more complex models: Consistency of effects and consistency in similar phases

The illustrations presented in the text so far refer to the simplest model, in which only a mean difference is modeled in the absence of trend. In the current section, we present the results for a model that also includes general trend and change in trend after introducing the intervention. For such a model, it is most common to code and interpret the change in level as an immediate change taking place during the first intervention phase measurement occasion (Moeyaert et al.,2014b). Moreover, there are two effects whose consistency can be assessed: the immediate change in level and the change in trend. Additionally, it is also possible to perform a more complete evaluation of the consistency of data in similar phases by comparing the intercept (initial baseline level) and the baseline trend across participants. In contrast, in the simpler models presented previously, for performing an assessment of the consistency of similar phases we could have focused only on the intercept, which then would have represented the average baseline level.

The more complex model can be applied to the Lambert et al. (2006) data, as Moeyaert et al. (2014a) and Shadish et al. (2014) also discuss possible baseline trends. We refer to this model as "Model.L3": Table 2 includes the numerical results for the fixed effect (baseline level, immediate change in level, baseline trend, and change in trend) and the standard deviations representing the random effects. Additionally, the individual empirical Bayes estimates were as follows: (a) for baseline level, ranging from 5.65 to 6.54; (b) for immediate change in level, ranging from −8.65 to −3.89; (c) for baseline trend, ranging from −0.05 to 0.25; and (d) for change in trend, ranging from −0.98 to 1.07. The graphical representation of the model is shown in Fig. 7.

For assessing consistency, the caterpillar plot for the A_1-B_1 comparison is presented in Fig. 8. In terms of consistency of effect, all nine confidence intervals include the fixed effect estimate for the immediate change in level, whereas for the change in trend, eight of the nine confidence intervals include the fixed effect estimate. According to the coefficient of variation, for the immediate change in level, there is very small variability and high consistency: $100 \times (0.423604/|-6.1857317|) = 6.85\%$. Given that the estimate for the change in trend is very close to zero (i.e., there is practically no change in trend), the coefficient of variation suggests less consistency ($100 \times (0.575299/|-0.2671592|) = 215.34\%$), but it should not be the main quantification for such a small effect. In terms of consistency of data in similar phases, focusing on the baseline, eight of the nine confidence intervals include the fixed effect estimate for the intercept and for the baseline trend. The coefficient of variation for the intercept is $100 \times (0.424273/|6.2486904|) = 6.79\%$, whereas that for the baseline trend is $100 \times (0.081208/|0.1432806|) = 56.68\%$. Once again, there is apparently lower consistency in the baseline trend, but this is related to the data presenting almost no baseline trend on average.

Visual inspection of the Sherer and Schreibman (2005) data suggests that there are different trends in the baseline and intervention phase, which makes the more complex model

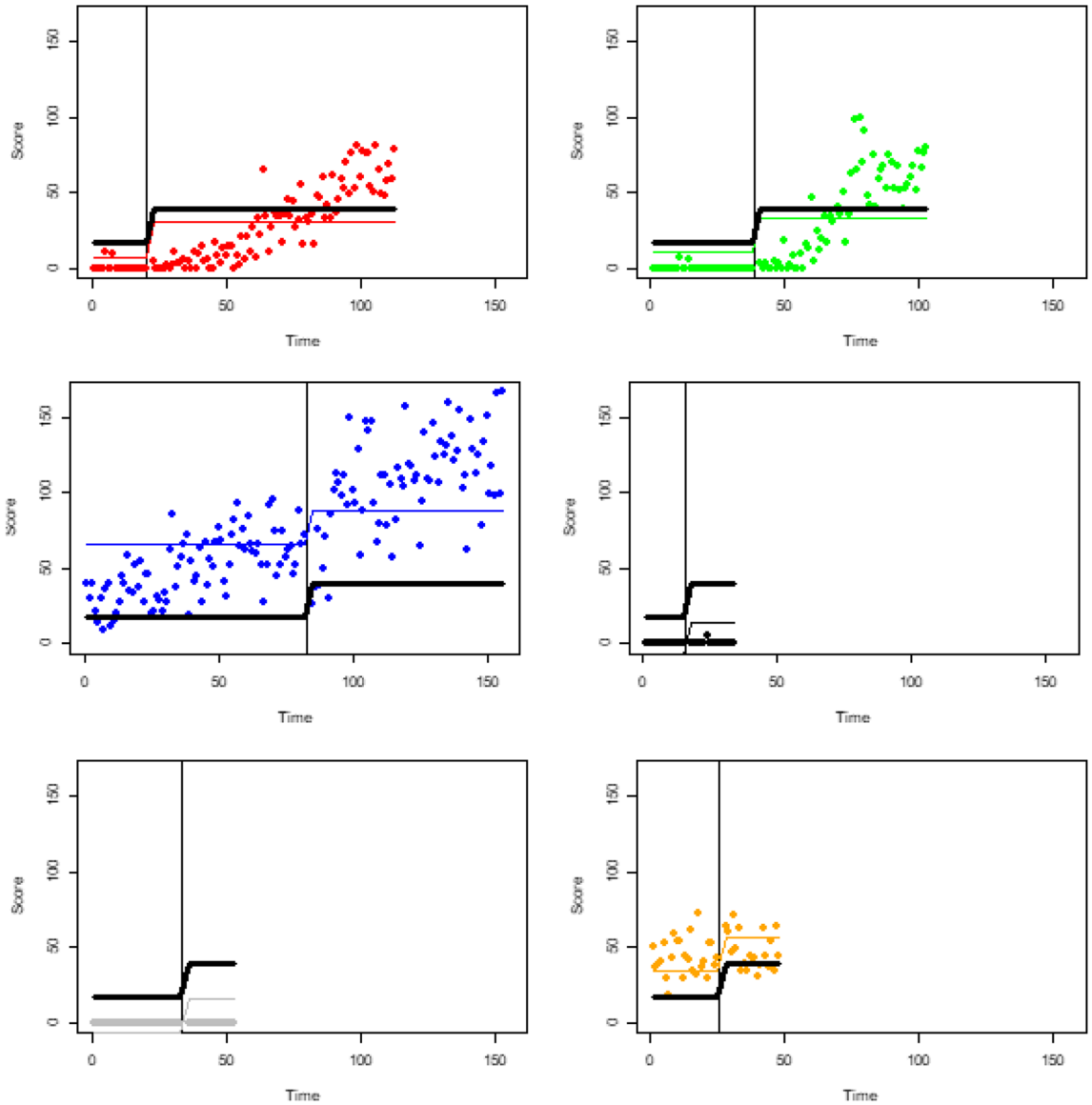

Fig. 5 Graphical representation of the multilevel model representing change in level, applied to the Sherer and Schreibman (2005) data

reasonable. We refer to this model as "Model.S2": Table 2 includes the numerical results for fixed effect (baseline level, immediate change in level, baseline trend, and change in trend) and the standard deviations representing the random effects. Additionally, the individual empirical Bayes estimates were as follows: (a) for baseline level, ranging from −1.71 to 37.46; (b) for immediate change in level, ranging from −5.63 to 11.55; (c) for baseline trend, ranging from −0.08 to 0.57; and (d) for change in trend, ranging from −0.68 to 1.27. The graphical representation of the model is shown in Fig. 9.

The caterpillar plot is presented in Fig. 10. Regarding the consistency of effects, none of the six confidence intervals includes the fixed effect estimate for the immediate change in level, whereas for the change in trend, two of the six confidence intervals include the fixed effect estimate. Accordingly, the coefficient of variation is very high in both cases: $100 \times (4.55251/|2.584279|) = 176.16\%$ for the immediate change in level and $100 \times (0.71328/|0.552067|) = 129.20\%$ for the change in trend. For the consistency of the baseline phases, none of the confidence intervals includes the fixed effect estimate of the intercept, and only one includes the fixed effect estimate for the baseline trend. Accordingly, the coefficient of variation is very high in both cases: $100 \times (16.85489/|10.99686|) = 153.27\%$ for the intercept and $100 \times (0.28957/|0.12886|) = 224.71\%$ for the baseline trend.

In summary, the graphical representation of the confidence intervals for the random effects can be used to distinguish between a data set with more consistent and successful

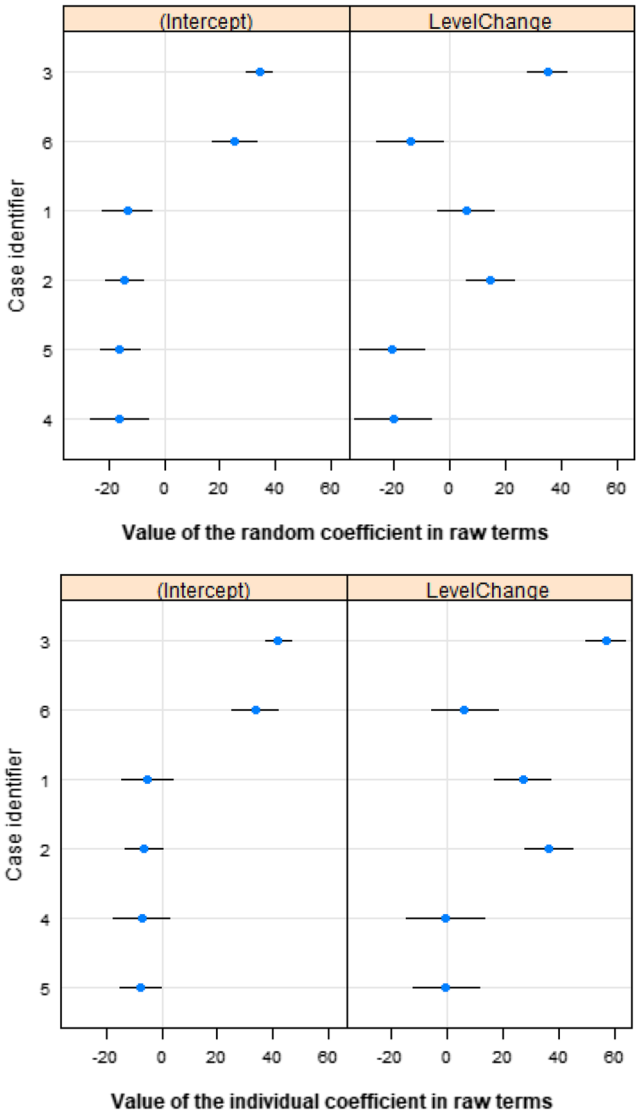

Fig. 6 The upper panel includes the caterpillar plot for the random effects and their confidence intervals, as obtained via a multilevel model including only change in level, for the Sherer and Schreibman (2005) data. The lower panel includes the empirical Bayes estimates of the individual treatment effects

replications (Lambert et al.) and a data set with lower consistency in similar phases and lower consistency in effects (Sherer and Schreibman).

Beyond multiple-baseline designs

Multilevel modeling in general, and the current proposals for assessing consistency of effects at the within-study level, is most straightforward for multiple-baseline designs. Actually, several reviews of published SCED research suggest that multiple-baseline designs are the most commonly used designs (Hammond & Gast, 2010; Shadish & Sullivan, 2011; Smith, 2012), present in more than half of the articles reviewed.

For other SCEDs, some decisions need to be made before applying a multilevel model. For instance, for an across-participant replicated ABAB, several design matrices are possible, allowing for different comparisons (Moeyaert et al., 2014b). For an across-participant replicated alternating treatments design, if there is an initial baseline phase before the comparison phase with rapid alternation of conditions, it is possible to compare the baseline to each of the alternating conditions (Moeyaert et al., 2014b). Otherwise, the average difference between the alternating conditions can be computed (Shadish et al., 2013). For a changing criterion design, one option is to compare the baseline phase to the last intervention subphase, i.e., for the final criterion level (Faith et al., 1996). Another option is to quantify the slope of the trend line across all intervention subphases (Shadish et al., 2013). It should be noted that for applying a multilevel model and for assessing the consistency of effects within a study, it is necessary to replicate the reversal/withdrawal, alternating treatments, or changing criterion design across participants. Once the appropriate design matrix is constructed and the multilevel analysis is carried out, the assessment of the consistency of effects can be performed as described in the previously presented examples.

Discussion

The research on multilevel models in their application to a single study has primarily focused on studying the estimation of fixed and random effects, as well as the coverage of confidence intervals (e.g., Baek & Ferron, 2013; Ferron et al., 2009; Ferron et al., 2010; Ferron et al., 2014; Moeyaert et al., 2017), type I error and power (Heyvaert et al., 2017), or dealing with count data (Declercq et al., 2019). Thus, the focus of the current text (namely, consistency of effects) is novel and it complements previous research. Moreover, the focus on consistency is well aligned with recent research on the topic (Tanious, De, Michiels, et al., 2019b; Tanious, Manolov et al., 2019c). As a strength of the proposal made here, this assessment of consistency can be performed using a free user-friendly website, and it can be easily represented visually. This makes it more likely to be accepted by applied researchers.

The assessment of consistency in the context of model building

One of the questionable research practices mentioned in relation to the "replicability crisis" is the ambiguous choices regarding data analysis (Hantula, 2019), which could be countered by preregistering analysis plans (Hales et al., 2019). A

Table 2. Results from applying multilevel models representing immediate change in level and change in trend, using the lme4 package

Aspect	Average estimate	Standard error	Standard deviation
Model.L3 for the Lambert et al. (2006) data			
Baseline level	6.23	0.52	0.34
Immediate change in level	−7.05	0.09	1.66
Baseline trend	0.15	0.94	0.11
Change in trend	−0.28	0.31	0.71
Model.S2 for the Sherer and Schreibman (2005) data			
Baseline level	10.92	8.93	20.07
Immediate change in level	4.93	8.29	6.89
Baseline trend	0.14	0.29	0.31
Change in trend	0.55	0.48	0.83

Note. The average estimate represents the fixed effect, whereas the standard deviation represents the random effect

multilevel model, just like the BC-SMD (Shadish et al., 2014), imposes the same kind of quantification for all participants for whom the different conditions are being compared. Such an analytical practice avoids the possibility of adapting the analysis or the quantitative emphasis to the most salient features of the data. However, the flexibility of multilevel models comes with the price of many decisions that need to be made regarding the exact model to apply (Baek et al., 2016).

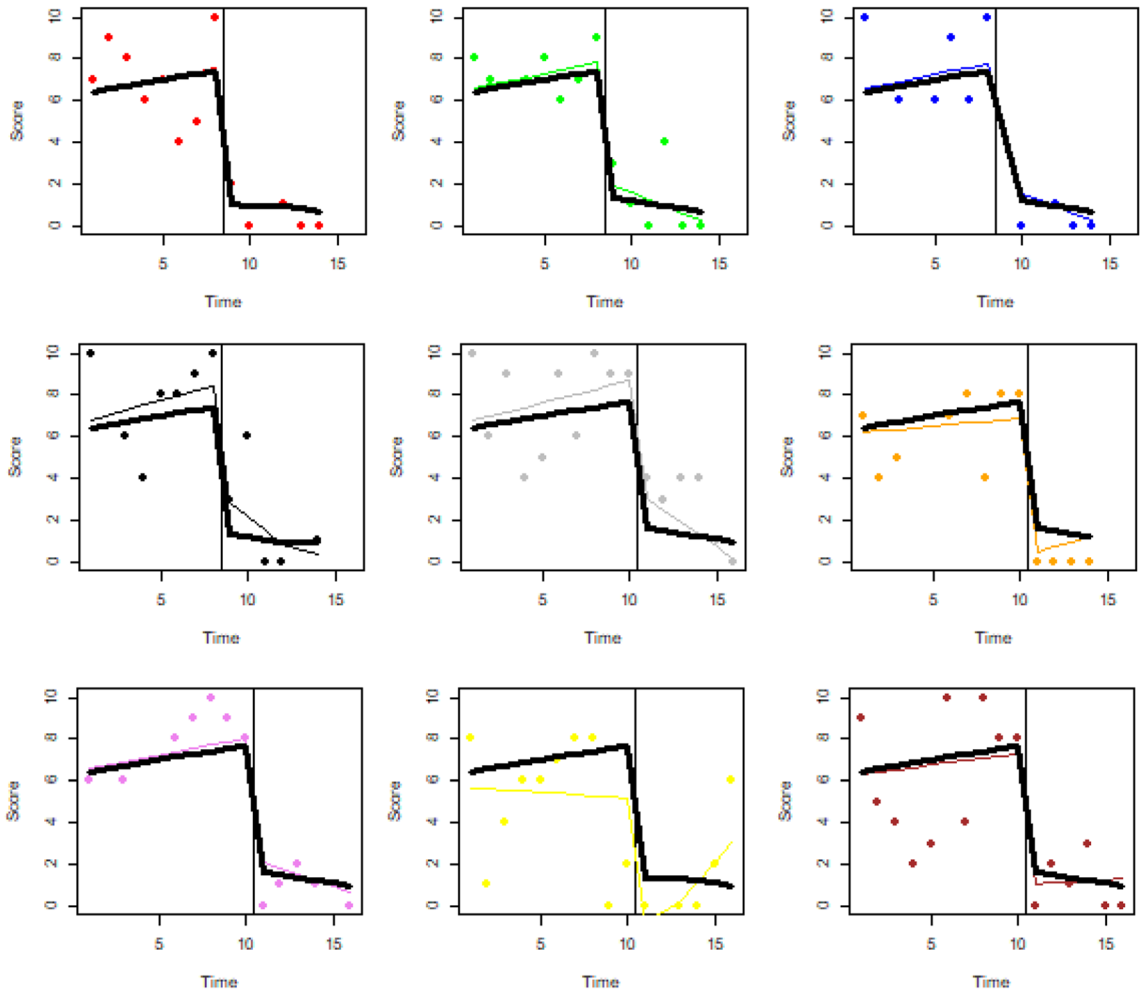

Fig. 7. Graphical representation of the multilevel model representing immediate change in level and change in slope, applied to the A_1-B_1 comparisons from the Lambert et al. (2006) data

In relation to model building, the decisions (e.g., include trend or not, which effects to include as random) could be made in relation to what is visible on the plots of raw data, but such a practice potentially leads to overfitting (Baek et al., 2016; Hox, 2010). The subjective visual inspection can be complemented by fit indices (e.g., the Akaike or the Bayesian information criterion) for deciding whether a more complex model offers sufficient improvement in fit (Dedrick et al., 2009; Ferron et al., 2008). In that sense, more complex models are not necessarily desirable, given that they may entail estimation problems and require larger samples (Wiley & Rapp, 2019). A different kind of comparison across models can be made via sensitivity analysis: checking the degree to which the conclusions change for different modeling options (Baek & Ferron, 2013; Moeyaert et al., 2014a).

In order to avoid excessively data-driven decisions and to reduce the possibility of overfitting, the model can be selected prior to data collection on the basis of theoretical considerations and previous evidence (Ferron et al., 2008; Onghena et al., 2018; Wiley & Rapp, 2019). For instance, the decision of whether to model baseline trend can be based on the expectations regarding spontaneous improvement (e.g., in neurorehabilitation, Krasny-Pacini & Evans, 2018) or on knowledge about baseline stability (Baek et al., 2014), whereas the choice of whether to model change in trend can be related to whether a gradual effect is expected (e.g., in academic interventions, Maggin et al., 2018). Additionally, if modeling trend is considered necessary, Shadish et al. (2013) suggest that both random intercepts and random slopes are needed for the proper modeling of autocorrelation. In fact, several illustrations of the use of multilevel models incorporating terms for trend include both random intercepts and random slopes (e.g., Baek et al., 2014; Gage & Lewis, 2014; Moeyaert et al., 2014a). In the context of the current proposal, the inclusion of random intercepts and random slopes allows for the assessment of consistency in similar phases (i.e., consistency of baseline level and baseline trend across cases) and consistency of effects (i.e., consistency of change in level and change in slope.

In summary, considering that all models are wrong (Box & Draper, 1987), trying out multiple models without an a priori basis may lead not only to capitalizing on chance, but also to ethical concerns (Levin et al., 2017). Thus, we recommend that the fundament for the model chosen should be at least partially related to the expectations stemming from the available literature, whereas visual analysis can still be used post hoc in order to comment on the meaningfulness of these quantifications (Parker et al., 2006). Specifically in relation to the

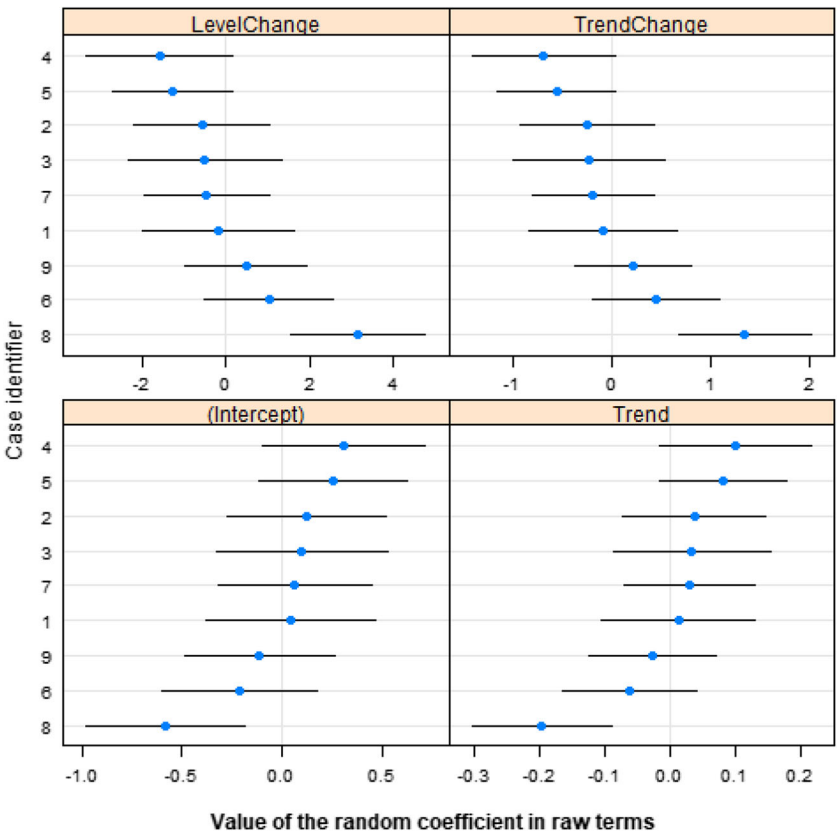

Fig. 8 Caterpillar plot for the random effects of a multilevel model including trend, change in trend, and immediate change in level, for the A_1-B_1 comparison from Lambert et al. (2006)

current proposal, including a predefined criterion for what is considered a successful and consistent replication, as suggested here, is expected to lead to results that are less affected by the "researcher degrees of freedom" (Hantula, 2019).

Software considerations

For the proposals made in the current text, we opted for a software implementation in R because it offers the possibility to create a freely available menu-driven website, via the Shiny package. In contrast, software such as SAS (which has been previously presented for using multilevel models; Baek & Ferron, 2013; Ferron et al., 2014; Moeyaert et al., 2014a; Moeyaert et al., 2013) is commercial and would require that the user works with programming code (syntax).

Using the website https://manolov.shinyapps.io/ExpectedPattern/ it is possible to obtain both numerical results and graphical representations. The website provides an example of the expected data structure, whereas the example data sets used in the current text can be obtained from https://osf.io/p3bna/. Once a data file is located and loaded, it is possible to specify expectations, such as the presence of baseline trend or the immediacy of effect, that help in choosing a multilevel model. After the expectations are specified, the quantitative results of multilevel models are obtained, along with line graphs representing the measurements for all participants, with superimposed mean or trend lines. Additionally, caterpillar plots such as those included in the present text are also obtained.

However, given that the topic is consistency, we have to mention that there may be inconsistencies between the

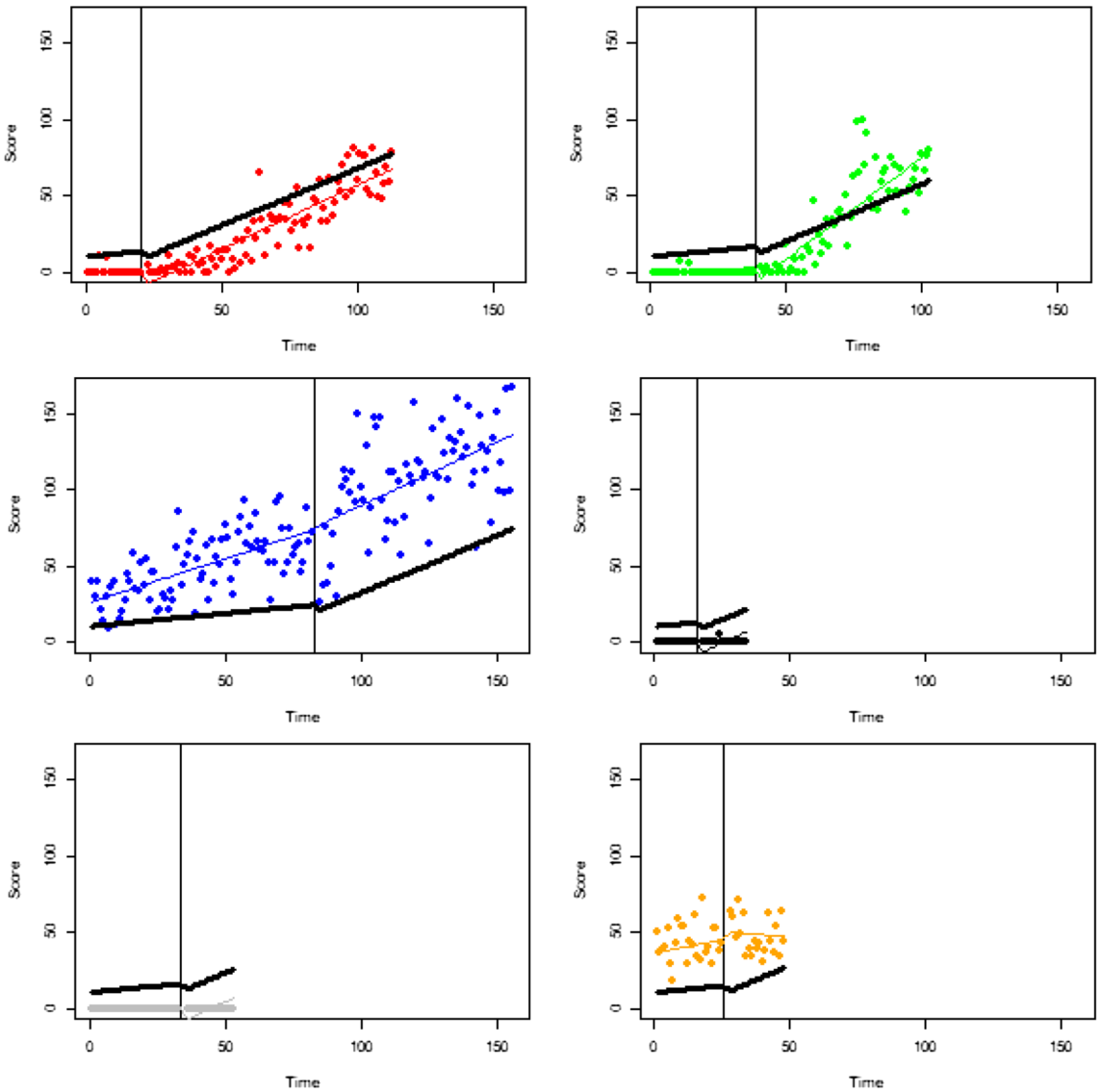

Fig. 9 Graphical representation of the multilevel model representing immediate change in level and change in slope, applied to the Sherer and Schreibman (2005) data

different software programs used for carrying out multilevel models, and even between different packages within R. Specifically, the *nlme* package (https://cran.r-project.org/web/packages/nlme/index.html) allows modelling for autocorrelation, which can be considered an advantage given the evidence available on the presence of autocorrelation in SCED data (Shadish & Sullivan, 2011). In contrast, the *lme4* package (https://cran.r-project.org/web/packages/lme4/index.html) does not offer this option, but has two advantages as compared with the *nlme* package: (a) the possibility to use the Kenward-Roger correction for degrees of freedom when obtaining *p* values (Wiley & Rapp, 2019), and (b) the automatic construction of caterpillar plots. In relation to the topic of the current text, to the best of our knowledge, caterpillar plots cannot be easily constructed for the objects resulting from a applying a multilevel model via the *nlme* package. Therefore, an optimal solution (modelling autocorrelation, using Kenward-Roger degrees of freedom, and obtaining caterpillar plots) is apparently currently not possible in R. Thus, the caterpillar plots obtained via the website https://manolov.shinyapps.io/ExpectedPattern/ are based on models without autocorrelation. For the sake of completeness and comparability, the numerical results from both the *nlme* and the *lme4* packages are included in the website.

Alternatively, software like SAS could be used to estimate the multilevel models, which would facilitate construction of confidence intervals that reflect Kenward-Roger adjusted degrees of freedom, but would require additional work to construct the caterpillar plots. Finally, it should be noted that the results obtained with other pieces of software for multilevel models, such as SAS PROC MIXED, MLwiN, SPSS Mixed, or HLM, cannot be expected to be completely identical. Therefore, despite the current developments and the software availability, more work is necessary for making all analytical options available across several instances of software in order to avoid suboptimal analyses.

Limitations and future research

The focus of the current text is on the quantification and graphical representation of the consistency of effects (i.e., direct replication) within a SCED study. Therefore, the illustrations are presented with verbal descriptions of the model building process. The reader interested in multilevel model building and formal presentation of the multilevel models within a single study, can refer to Baek et al.

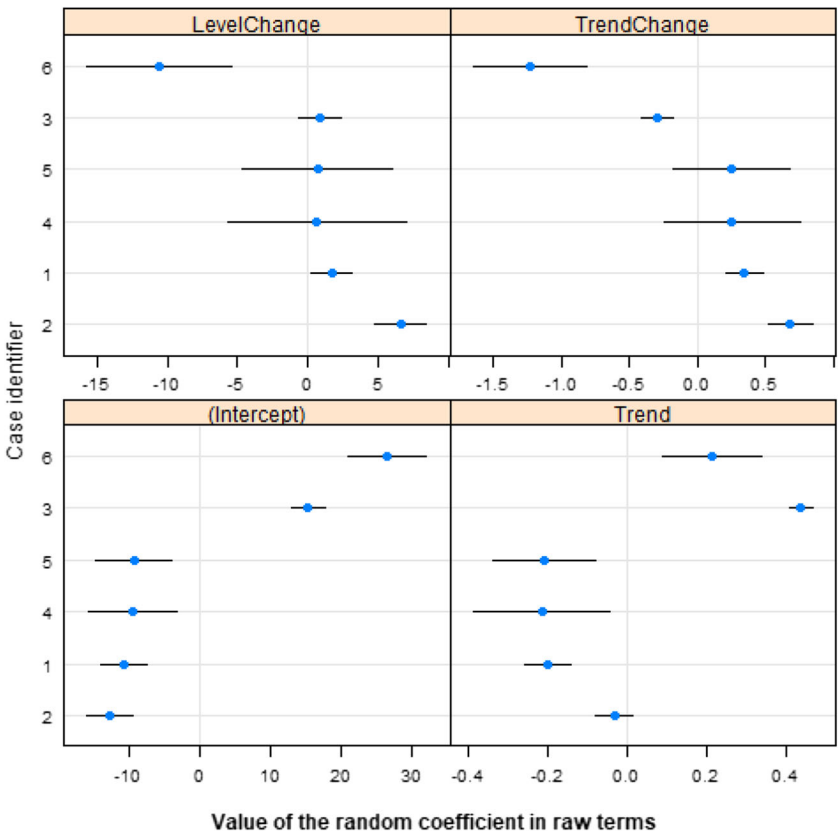

Fig. 10 Caterpillar plot for the random effects of a multilevel model including trend, change in trend, and immediate change in level, for the Sherer and Schreibman (2005) data

(2014), Dedrick et al. (2009), and Moeyaert, et al. (2014a).

Provided that the focus is on within-study replication, we did not deal extensively with replication across studies. Although certain uses of SCEDs are not aimed at demonstrating the generality of the intervention effects (Riley-Tillman & Burns, 2009), if the aim is to establish the generalizability of the intervention effects, systematic replications across studies are relevant (Maggin, 2015; Onghena et al., 2018; Tate & Perdices, 2019). Even when generalization is desirable, the external validity in the SCED context is not an issue of statistical inference and extrapolation, but rather follows a more inductive approach (Kennedy, 2005). In this approach, the descriptions of participants, interventions, target behaviors, and settings are crucial (Maggin, 2015; Tate et al., 2013), and the amount of generality can be understood as a continuum according to the number of variables (related to participants, target behaviors, and settings) that change in systematic replications across studies (Gast & Ledford, 2018; Riley-Tillman & Burns, 2009). In that sense, failing to replicate an effect allows one to discover the limitations of an intervention, which is also useful for prompting further modifications and further research for better understanding the reasons an intervention does or does not work (Gast & Ledford, 2018). Finally, building the evidence about generality on the basis of a series of individual studies makes the meta-analyses of the SCED studies and multilevel models relevant (Jenson et al., 2007; Onghena et al., 2018).

Regarding the proposal of quantifying the percentage of random effect confidence intervals that include 0, it should be noted that this percentage is not expected to approximate any theoretically desirable quantity. In that sense, we are not quantifying how many of the confidence intervals in different samples or replications include a population parameter, which would be equivalent to studying the coverage of a confidence interval (e.g., Baek et al., 2019; Ferron et al., 2009; Moeyaert et al., 2017), expected to be .95 for a 95% confidence. Additionally, what we are proposing is not the same as estimating the capture percentage of an initial confidence interval in reference to the means of subsequent replications, expected to be equal to .83 for a 95% confidence (called "prediction interval for a replication mean" by Cumming, 2012). Therefore, for the percentage of random effects confidence intervals that include 0, there is not an exact cutoff point that suggests sufficient consistency, just like experimental control should be understood as a continuum and not as something that is either present or absent (Horner & Odom, 2014). The 3:1 ratio (Maggin et al., 2013) and the corresponding percentage of 75% is only an indication, and not a fixed criterion. Nevertheless, it has been highlighted that statistical thinking is more important than mechanically applying a given ritual (Gigerenzer, 2004).

In terms of the statistical properties of the quantifications proposed, some comments are necessary. There is evidence that the confidence intervals for the variance of the treatment effect (i.e., change in level) present undercovering (Ferron et al., 2009). However, it is unclear whether this evidence can be extrapolated to each of the confidence intervals for the random effects (i.e., for the confidence intervals for the difference between the individual treatment effects and the fixed effect estimate). Similarly, it is not clear whether the evidence about the confidence intervals for the individual treatment effects (wider intervals, but better coverage when using the Kenward-Roger estimation of the degrees of freedom; Ferron et al., 2010) can be extrapolated to the confidence intervals for the difference between the individual treatment effects and the fixed effect estimate. Therefore, more research is needed on the latter.

Open Practices Statements

The current text is not based on gathering data (e.g., in the context of an experiment). Therefore, there are no primary data or materials to be made available and there is no empirical study requiring preregistration. Nonetheless, the data used for the illustrations and the R code for constructing the caterpillar plots are available at https://osf.io/p3bna/.

References

Baek, E. K., & Ferron, J. M. (2013). Multilevel models for multiple-baseline data: Modeling across-participant variation in autocorrelation and residual variance. *Behavior Research Methods, 45*(1), 65–74. https://doi.org/10.3758/s13428-012-0231-z

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

Baek, E. K., Petit-Bois, M., Van den Noortgate, W., Beretvas, S. N., & Ferron, J. M. (2016). Using visual analysis to evaluate and refine multilevel models of single-case studies. *The Journal of Special Education, 50*(1), 18–26. https://doi.org/10.1177/0022466914565367

Baek, E. K., Beretvas, S. N., Van den Noortgate, W., & Ferron, J. M. (2019). Brief research report: Bayesian versus REML estimations with noninformative priors in multilevel single-case data. *The Journal of Experimental Education*. Advance online publication. - https://doi.org/10.1080/00220973.2018.1527280

Barker, J., McCarthy, P., Jones, M., & Moran, A. (2011). *Single case research methods in sport and exercise psychology*. Routlege.

Barlow, D., Nock, M., & Hersen, M. (2009). *Single case experimental designs: Strategies for studying behavior change* (3rd). Allyn and Bacon.

Borenstein, M., Higgins, J., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. *Research Synthesis Methods, 8*(1), 5–18. https://doi.org/10.1002/jrsm.1230

Box, G. E. P., & Draper, N. R. (1987). *Empirical model-building and response surfaces*. Wiley.

Branch, M. N. (2019). The "reproducibility crisis:" Might the methods used frequently in behavior-analysis research help? *Perspectives on Behavior Science, 42*(1), 77–89. https://doi.org/10.1007/s40614-018-0158-5

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Cook, B. G., Buysse, V., Klingner, J., Landrum, T. J., McWilliam, R. A., Tankersley, M., & Test, D. W. (2015). CEC's standards for classifying the evidence base of practices in special education. *Remedial and Special Education, 36*(4), 220–234. https://doi.org/10.1177/0741932514557271

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Taylor & Francis.

Davis, D. H., Gagné, P., Fredrick, L. D., Alberto, P. A., Waugh, R. E., & Haardörfer, R. (2013). Augmenting visual analysis in single-case research with hierarchical linear modeling. *Behavior Modification, 37*(1), 62–89. https://doi.org/10.1177/0145445512453734

Declercq, L., Jamshidi, L., Fernández-Castilla, B., Beretvas, S. N., Moeyaert, M., Ferron, J. M., & Van den Noortgate, W. (2019). Analysis of single-case experimental count data using the linear mixed effects model: A simulation study. *Behavior Research Methods, 51*(6), 2477-2497. https://doi.org/10.3758/s13428-018-1091-y

Dedrick, R. F., Ferron, J. M., Hess, M. R., Hogarty, K. Y., Kromrey, J. D., Lang, T. R., Niles, J. D., & Lee, R. S. (2009). Multilevel modeling: A review of methodological issues and applications. *Review of Educational Research, 79*(1), 69–102. https://doi.org/10.3102/0034654308325581

DeHart, W. B., & Kaplan, B. A. (2019). Applying mixed-effects modeling to single-subject designs: An introduction. *Journal of the Experimental Analysis of Behavior, 111*(2), 192–206. https://doi.org/10.1002/jeab.507

Dixon, M. A., & Cunningham, G. B. (2006). Data aggregation in multilevel analysis: A review of conceptual and statistical issues. *Measurement in Physical Education and Exercise Science, 10*(2), 85–107. https://doi.org/10.1207/s15327841mpee1002_2

Edgington, E. S., & Onghena, P. (2007). *Randomization tests* (4th). Chapman & Hall/CRC.

Faith, M. S., Allison, D. B., & Gorman, D. B. (1996). Meta-analysis of single-case research. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 245–277). Lawrence Erlbaum Associates.

Ferron, J. M., Hogarty, K. Y., Dedrick, R. F., Hess, M. R., Niles, J. D., & Kromrey, J. D. (2008). Reporting results from multilevel analyses. In A. A. O'Connell and D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 391–426). Information Age Publishing.

Ferron, J. M., Bell, B. A., Hess, M. R., Rendina-Gobioff, G., & Hibbard, S. T. (2009). Making treatment effect inferences from multiple-baseline data: The utility of multilevel modeling approaches. *Behavior Research Methods, 41*(2), 372–384. https://doi.org/10.3758/BRM.41.2.372

Ferron, J. M., Farmer, J. L., & Owens, C. M. (2010). Estimating individual treatment effects from multiple-baseline data: A Monte Carlo study for multilevel-modeling approaches. *Behavior Research Methods, 42*(4), 930–943. https://doi.org/10.3758/BRM.42.4.930

Ferron, J. M., Moeyaert, M., Van den Noortgate, W., & Beretvas, S. N. (2014). Estimating causal effects from multiple-baseline studies: Implications for design and analysis. *Psychological Methods, 19*(4), 493–510. https://doi.org/10.1037/a0037038

Gage, N. A., & Lewis, T. J. (2014). Hierarchical linear modeling meta-analysis of single-subject design research. *The Journal of Special Education, 48*(1), 3–16. https://doi.org/10.1177/0022466912443894

Ganz, J. B., & Ayres, K. M. (2018). Methodological standards in single-case experimental design: Raising the bar. *Research in Developmental Disabilities, 79*(1), 3–9. https://doi.org/10.1016/j.ridd.2018.03.003

Garrett, H. E. (1937). *Statistics in psychology and education* (2nd). Oxford, UK: Longmans, Green.

Gast, D. L., & Ledford, J. R. (2018). Replication. In J. R. Ledford & D. L. Gast (Eds.), *Single case research methodology: Applications in special education and behavioral sciences* (3rd ed.) (pp. 77–96). Routledge.

Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics, 33*(5), 587–606. https://doi.org/10.1016/j.socec.2004.09.033

Goh, J. X., Hall, J. A., & Rosenthal, R. (2016). Mini meta-analysis of your own studies: Some arguments on why and a primer on how. *Social & Personality Psychology Compass, 10*(10), 535–549. https://doi.org/10.1111/spc3.12267

Hagopian, L. P. (2020). The consecutive controlled case series: Design, data-analytics, and reporting methods supporting the study of generality. *Journal of Applied Behavior Analysis, 53*(2), 596-619. https://doi.org/10.1002/jaba.691

Hales, A. H., Wesselmann, E. D., & Hilgard, J. (2019). Improving psychological science through transparency and openness: An overview. *Perspectives on Behavior Science, 42*(1), 13–31. https://doi.org/10.1007/s40614-018-00186-8

Hammond, D., & Gast, D. L. (2010). Descriptive analysis of single subject research designs: 1983-2007. *Education and Training in Autism and Developmental Disabilities, 45*(2), 187–202. https://www.jstor.org/stable/23879806

Hantula, D. A. (2019). Editorial: Replication and reliability in behavior science and behavior analysis: A call for a conversation. *Perspectives on Behavior Science, 42*(1), 1–11. https://doi.org/10.1007/s40614-019-00194-2

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2012). A standardized mean difference effect size for single case designs. *Research Synthesis Methods, 3*(3), 224−239. https://doi.org/10.1002/jrsm.1052

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2013). A standardized mean difference effect size for multiple baseline designs across individuals. *Research Synthesis Methods, 4*(4), 324−341. https://doi.org/10.1002/jrsm.1086

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

Horner, R. H., & Kratochwill, T. R. (2012). Synthesizing single-case research to identify evidence-based practices: Some brief reflections. *Journal of Behavioral Education, 21*(3), 266−272. https://doi.org/10.1007/s10864-012-9152-2

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children, 71*(2), 165−179. https://doi.org/10.1177/001440290507100203

Horner, R. J., & Odom, S. L. (2014). Constructing single-case research designs: Logic and options. In T. R. Kratochwill & J. R. Levind (Eds.), *Single-case intervention research: Methodological and statistical advances* (pp. 27–51). American Psychological Association. https://doi.org/10.1037/14376-002

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

Janosky, J. E., Leininger, S. L., Hoerger, M. P., & Libkuman, T. M. (2009). *Single subject designs in biomedicine*. Springer.

Jenson, W. R., Clark, E., Kircher, J. C., & Kristjansson, S. D. (2007). Statistical reform: Evidence-based practice, meta-analyses, and single subject designs. *Psychology in the Schools, 44*(5), 483−493. https://doi.org/10.1002/pits.20240

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

Killeen, P. R. (2019). Predict, control, and replicate to understand: How statistics can foster the fundamental goals of science. *Perspectives on Behavior Science, 42*(1), 109–132. https://doi.org/10.1007/s40614-018-0171-8

Klaassen, F. (2020). Combining evidence over multiple individual cases. In R. van de Schoot & M. Miočević (Eds.), *Small sample size solutions: A guide for applied researchers and practitioners* (pp. 126–138). Routledge.

Krasny-Pacini, A., & Evans, J. (2018). Single-case experimental designs to assess intervention effectiveness in rehabilitation: A practical guide. *Annals of Physical and Rehabilitation Medicine, 61*(3), 164–179. https://doi.org/10.1016/j.rehab.2017.12.002

Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods, 15*(2), 124−144. https://doi.org/10.1037/a0017736

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

Kratochwill, T. R., Levin, J. R., & Horner, R. H. (2018). Negative results: Conceptual and methodological dimensions in single-case intervention research. *Remedial and Special Education, 34*(1), 26–38. https://doi.org/10.1177/0741932512452794

Lambert, M. C., Cartledge, G., Heward, W. L., & Lo, Y. (2006). Effects of response cards on disruptive behavior and academic responding during math lessons by fourth-grade urban students. *Journal of Positive Behavior Interventions, 8*(2), 88–99. https://doi.org/10.1177/10983007060080020701

Lane, J. D., Ledford, J. R., & Gast, D. L. (2017). Single-case experimental design: current standards and applications in occupational therapy. *American Journal of Occupational Therapy, 71*(2), 7102300010p1–7102300010p9. https://doi.org/10.5014/ajot.2017.022210

Lanovaz, M. J., Turgeon, S., Cardinal, P., & Wheatley, T. L. (2019). Using single-case designs in practical settings: Is within-subject replication always necessary? *Perspectives on Behavior Science, 42*(1), 153–162. https://doi.org/10.1007/s40614-018-0138-9

Ledford, J. R. (2018). No randomization? No problem: Experimental control and random assignment in single case research. *American Journal of Evaluation, 39*(1), 71–90. https://doi.org/10.1177/1098214017723110

Ledford, J. R., & Gast, D. L. (Eds.) (2018). *Single case research methodology: Applications in special education and behavioral sciences* (3rd). Routledge.

Ledford, J. R., Barton, E. E., Hardy, J. K., Elam, K., Seabolt, J., Shanks, M., Hemmeter, M. L., & Kaiser, A. (2016). What equivocal data from single case comparison studies reveal about evidence-based practices in early childhood special education. *Journal of Early Intervention, 38*(2), 79−91. https://doi.org/10.1177/1053815116648000

Ledford, J. R., Barton, E. E., Severini, K. E., & Zimmerman, K. N. (2019). A primer on single-case research designs: Contemporary use and analysis. *American Journal on Intellectual and Developmental Disabilities, 124*(1), 35–56. https://doi.org/10.1352/1944-7558-124.1.35

Levin, J. R., Ferron, J. M., & Gafurov, B. S. (2017). Additional comparisons of randomization-test procedures for single-case multiple-baseline designs: Alternative effect types. *Journal of School Psychology, 63*, 13–34. https://doi.org/10.1016/j.jsp.2017.02.003

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

Maggin, D. M. (2015). Considering generality in the systematic review and meta-analysis of single-case research: A response to Hitchcock et al. *Journal of Behavioral Education, 24*(4), 470-482. https://doi.org/10.1007/s10864-015-9239-7

Maggin, D. M., Briesch, A. M., & Chafouleas, S. M. (2013). An application of the What Works Clearinghouse standards for evaluating single-subject research: Synthesis of the self-management literature base. *Remedial and Special Education, 34*(1), 44–58. https://doi.org/10.1177/0741932511435176

Maggin, D. M., Lane, K. L., & Pustejovsky, J. E. (2017). Introduction to the special issue on single-case systematic reviews and meta-analyses. *Remedial and Special Education, 38*(6), 323−330. https://doi.org/10.1177/0741932517717043

Maggin, D. M., Cook, B. G., & Cook, L. (2018). Using single-case research designs to examine the effects of interventions in special education. *Learning Disabilities Research & Practice, 33*(4), 182–191. https://doi.org/10.1111/ldrp.12184

Manolov, R. (2018). Linear trend in single-case visual and quantitative analyses. *Behavior Modification, 42*(5), 684–706. https://doi.org/10.1177/0145445517726301

Manolov, R., & Moeyaert, M. (2017). Recommendations for choosing single-case data analytical techniques. *Behavior Therapy, 48*(1), 97–114. https://doi.org/10.1016/j.beth.2016.04.008

Michiels, B., & Onghena, P. (2019a). Randomized single-case AB phase designs: Prospects and pitfalls. *Behavior Research Methods, 51*(6), 2454–2476. https://doi.org/10.3758/s13428-018-1084-x

Michiels, B., & Onghena, P. (2019b). Nonparametric meta-analysis for single-case research: Confidence intervals for combined effect sizes. *Behavior Research Methods, 51*(3), 1145–1160. https://doi.org/10.3758/s13428-018-1044-5

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

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*(5), 719−748. https://doi.org/10.1080/00273171.2013.816621

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

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

Moeyaert, M., Rindskopf, D., Onghena, P., & Van den Noortgate, W. (2017). Multilevel modeling of single-case data: A comparison of maximum likelihood and Bayesian estimation. *Psychological Methods, 22*(4), 760–778. https://doi.org/10.1037/met0000136

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*(2), 241–301. https://doi.org/10.1037/1082-989X.5.2.241

Ninci, J. (2019). Single-case data analysis: A practitioner guide for accurate and reliable decisions. *Behavior Modification.* Advance online publication. https://doi.org/10.1177/0145445519867054

Nolte, K. F. (1937). Simplification of vocabulary and comprehension in reading. *The Elementary English Review, 14*(4), 119–146. https://www.jstor.org/stable/41380939

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

experiments. *Brain Impairment, 19*(1), 33–58. https://doi.org/10.1017/BrImp.2017.25

Parker, R. I., & Vannest, K. J. (2009). An improved effect size for single-case research: Nonoverlap of all pairs. *Behavior Therapy, 40*(4), 357–367. https://doi.org/10.1016/j.beth.2008.10.006

Parker, R. I., Cryer, J., & Byrns, G. (2006). Controlling baseline trend in single-case research. *School Psychology Quarterly, 21*(4), 418–443. https://doi.org/10.1037/h0084131

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

Peng, C. Y. J., & Chen, L. T. (2015). Algorithms for assessing intervention effects in single-case studies. *Journal of Modern Applied Statistical Methods, 14*(1), 276–307. https://doi.org/10.22237/jmasm/1430452800

Perdices, M., Tate, R. L., & Rosenkoetter, U. (2019). An algorithm to evaluate methodological rigor and risk of bias in single-case studies. *Behavior Modification.* Advance online publication. https://doi.org/10.1177/0145445519863035

Petursdottir, A. I., & Carr, J. E. (2018). Applying the taxonomy of validity threats from mainstream research design to single-case experiments in applied behavior analysis. *Behavior Analysis in Practice, 11*(3), 228-240. https://doi.org/10.1007/s40617-018-00294-6

Pustejovsky, J. E. (2018). Using response ratios for meta-analyzing single-case designs with behavioral outcomes. *Journal of School Psychology, 68*(Jun), 99−112. https://doi.org/10.1016/j.jsp.2018.02.003

Pustejovsky, J. E., Hedges, L. V., & Shadish, W. R. (2014). Design-comparable effect sizes in multiple baseline designs: A general modeling framework. *Journal of Educational and Behavioral Statistics, 39*(5), 368−393. https://doi.org/10.3102/1076998614547577

Riley-Tillman, T. C., & Burns, M. K. (2009). *Evaluating educational interventions: Single-case design for measuring response to intervention*. The Guilford Press.

Sanabria, F., & Killeen, P. R. (2007). Better statistics for better decisions: Rejecting null hypothesis statistical tests in favor of replication statistics. *Psychology in the Schools, 44*(5), 471–481. https://doi.org/10.1002/pits.20239

Schlosser, R. W. (2009). *The role of single-subject experimental designs in evidence-based practice times*. (FOCUS: Technical Brief 22). National Center for the Dissemination of Disability Research (NCDDR). Retrieved May 24, 2018 from http://ktdrr.org/ktlibrary/articles_pubs/ncddrwork/focus/focus22/Focus22.pdf

Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods, 43*(4), 971−980. https://doi.org/10.3758/s13428-011-0111-y

Shadish, W. R., Kyse, E. N., & Rindskopf, D. M. (2013). Analyzing data from single-case designs using multilevel models: New applications and some agenda items for future research. *Psychological Methods, 18*(3), 385–405. https://doi.org/10.1037/a0032964

Shadish, W. R., Hedges, L. V., & Pustejovsky, J. E. (2014). Analysis and meta-analysis of single-case designs with a standardized mean difference statistic: A primer and applications. *Journal of School Psychology, 52*(2), 123–147. https://doi.org/10.1016/j.jsp.2013.11.005

Shadish, W. R., Zelinsky, N. A. M., Vevea, J. L., & Kratochwill, T. R. (2016). A survey of publication practices of single-case design researchers when treatments have small or large effects. *Journal of Applied Behavior Analysis, 49*(3), 656−673. https://doi.org/10.1002/jaba.308

Sherer, M. R., & Schreibman, L. (2005). Individual behavioral profiles and predictors of treatment effectiveness for children with autism. *Journal of Consulting and Clinical Psychology, 73*(3), 525–538. https://doi.org/10.1037/0022-006X.73.3.525

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Smith, J. D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods, 17*(4), 510−550. https://doi.org/10.1037/a0029312

Swan, D. M., Pustejovsky, J. E., & Beretvas, S. N. (2020). The impact of response-guided designs on count outcomes in single-case experimental design baselines. *Evidence-Based Communication Assessment and Intervention, 14(1–2)*, 82–107. https://doi.org/10.1080/17489539.2020.1739048

Tanious, R., De, T. K., & Onghena, P. (2019a). A multiple randomization testing procedure for level, trend, variability, overlap, immediacy, and consistency in single-case phase designs. *Behaviour Research and Therapy, 119*(Aug), 103414. https://doi.org/10.1016/j.brat.2019.103414

Tanious, R., De, T. K., Michiels, B., Van den Noortgate, W., & Onghena, P. (2019b). Assessing consistency in single-case A-B-A-B phase designs. *Behavior Modification*. Advance online publication. https://doi.org/10.1177/0145445519837726

Tanious, R., Manolov, R., & Onghena, P. (2019c). The assessment of consistency in single-case experiments: Beyond A-B-A-B designs. *Behavior Modification.* Advance online publication. https://doi.org/10.1177/0145445519882889

Tate, R. L., & Perdices, M. (2019). *Single-case experimental designs for clinical research and neurorehabilitation settings: Planning, conduct, analysis, and reporting*. Routledge.

Tate, R. L., Perdices, M., Rosenkoetter, U., Wakim, D., Godbee, K., Togher, L., & McDonald, S. (2013). Revision of a method quality rating scale for single-case experimental designs and n-of-1 trials: The 15-item Risk of Bias in N-of-1 Trials (RoBiNT) Scale. *Neuropsychological Rehabilitation, 23*(5), 619–638. https://doi.org/10.1080/09602011.2013.824383

Tate, R. L., Perdices, M., Rosenkoetter, U., Shadish, W., Vohra, S., Barlow, D. H., Horner, R., Kazdin, A., Kratochwill, T. R., McDonald, S., Sampson, M., Shamseer, L., Togher, L., Albin, R., Backman, C., Douglas, J., Evans, J. J., Gast, D., Manolov, R., Mitchell, G., … , Wilson, B. (2016). The Single-Case Reporting guideline In BEhavioural interventions (SCRIBE) 2016 statement. *Journal of School Psychology, 56*, 133–142. https://doi.org/10.1016/j.jsp.2016.04.001

Tincani, M., & Travers, J. (2018). Publishing single-case research design studies that do not demonstrate experimental control. *Remedial and Special Education, 39*(2), 118–128. https://doi.org/10.1177/0741932517697447

Tincani, M., & Travers, J. (2019). Replication research, publication bias, and applied behavior analysis. *Perspectives on Behavior Science, 42*(1), 59–75. https://doi.org/10.1007/s40614-019-00191-5.

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods, 6*(4), 371–386. https://doi.org/10.1037/1082-989X.6.4.371

Van den Noortgate, W., & Onghena, P. (2003a). Combining single-case experimental studies using hierarchical linear models. *School Psychology Quarterly, 18*(3), 325–346. https://doi.org/10.1521/scpq.18.3.325.22577

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), 1–10. https://doi.org/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*(2), 196–209. https://doi.org/10.1037/h0100613

Wendt, O., & Miller, B. (2012). Quality appraisal of single-subject experimental designs: An overview and comparison of different appraisal tools. *Education and Treatment of Children, 35*(2), 235-268. https://doi.org/10.1353/etc.2012.0010

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

Wiley, R. W., & Rapp, B. (2019). Statistical analysis in Small-N Designs: using linear mixed-effects modeling for evaluating intervention effectiveness. *Aphasiology, 33*(1), 1–30. https://doi.org/10.1080/02687038.2018.1454884

Wolfe, K., Barton, E. E., & Meadan, H. (2019). Systematic protocols for the visual analysis of single-case research data. *Behavior Analysis in Practice*, *12*(2), 491–502. https://doi.org/10.1007/s40617-019-00336-7