Behavior Research Methods (2024) 56:379–405
https://doi.org/10.3758/s13428-022-02035-8

Examining the normality assumption of a design-comparable effect size in single-case designs

Li-Ting Chen[1] · Yi-Kai Chen[2] · Tong-Rong Yang[2] · Yu-Shan Chiang[3] · Cheng-Yu Hsieh[2,4] · Che Cheng[2] · Qi-Wen Ding[5] · Po-Ju Wu[6] · Chao-Ying Joanne Peng[2,6]

Accepted: 22 November 2022 / Published online: 17 January 2023

Abstract

What Works Clearinghouse (WWC, 2022) recommends a design-comparable effect size (D-CES; i.e., g_{AB}) to gauge an intervention in single-case experimental design (SCED) studies, or to synthesize findings in meta-analysis. So far, no research has examined g_{AB}'s performance under non-normal distributions. This study expanded Pustejovsky et al. (2014) to investigate the impact of data distributions, number of cases (*m*), number of measurements (*N*), within-case reliability or intra-class correlation (ρ), ratio of variance components (λ), and autocorrelation (ϕ) on g_{AB} in multiple-baseline (MB) design. The performance of g_{AB} was assessed by relative bias (*RB*), relative bias of variance (*RBV*), *MSE*, and coverage rate of 95% CIs (*CR*). Findings revealed that g_{AB} was unbiased even under non-normal distributions. g_{AB}'s variance was generally overestimated, and its 95% CI was over-covered, especially when distributions were normal or nearly normal combined with small *m* and *N*. Large imprecision of g_{AB} occurred when *m* was small and ρ was large. According to the ANOVA results, data distributions contributed to approximately 49% of variance in *RB* and 25% of variance in both *RBV* and *CR*. *m* and ρ each contributed to 34% of variance in *MSE*. We recommend g_{AB} for MB studies and meta-analysis with $N \geq 16$ and when either (1) data distributions are normal or nearly normal, $m = 6$, and $\rho = 0.6$ or 0.8, or (2) data distributions are mildly or moderately non-normal, $m \geq 4$, and $\rho = 0.2$, 0.4, or 0.6. The paper concludes with a discussion of g_{AB}'s applicability and design-comparability, and sound reporting practices of ES indices.

Keywords Single-case · Intervention · Standardized mean difference · Effect size · Design comparable · Normality

Single-case experimental designs (SCEDs) are research designs that can be used to determine whether there exists a causal or functional relationship between the introduction of an intervention and changes in outcome behavior(s). SCED studies have been used to evaluate the effectiveness of interventions in psychology, education, speech pathology, medicine, sports and athletic performance, to name a few (Barker et al., 2011; Byiers et al., 2012; Franklin et al., 1996; Horner et al., 2005; Kunze et al., 2021; Morgan & Morgan, 2009; Vlaeyen et al., 2020). SCED typically employs a small number of cases who serve as their own controls. Among the variety of SCEDs, the multiple baseline (MB) design was by far the most popular design accounting for nearly 50% of published SCED studies (Hammond & Gast, 2010; Horner & Odom, 2014; Pustejovsky et al., 2019; Shadish & Sullivan, 2011; Smith, 2012; Tanious & Onghena, 2021). An MB design consists of one A phase and one B phase across multiple cases, multiple behaviors of one case, or multiple settings for the same behavior of a case. During the A phase, a case (or a behavior) is observed to be stabilized before an intervention is introduced to that case (or that behavior). An intervention in an MB design is successively administered to all cases (behaviors or settings) until all cases (all behaviors, or a behavior in all settings) are intervened to allow for an assessment of its effectiveness.

We thank two reviewers and the Action Editor for their insightful comments that greatly improved this paper.

✉ Li-Ting Chen
litingc@unr.edu

[1] Department of Educational Studies, University of Nevada, Reno, Reno, NV, USA

[2] Department of Psychology, National Taiwan University, Taipei, Taiwan

[3] Department of Curriculum & Instruction, Indiana University Bloomington, Bloomington, IN, USA

[4] Department of Psychology, Royal Holloway, University of London, Egham, UK

[5] Institute of Sociology, Academia Sinica, Taipei, Taiwan

[6] Department of Counseling and Educational Psychology, Indiana University Bloomington, Bloomington, IN, USA

Advances in SCED methodology

Traditionally, systematic visual analysis of SCED data has been used to determine whether a behavioral change is due to the introduction of an intervention, but not chance fluctuations (Horner et al., 2005; Kazdin, 2011; Wolfe & McCammon, 2022). In recent years, advanced approaches have been applied to quantify intervention effects in SCED studies (Chen et al., 2019; Kazdin, 2019; Tanious & Manolov, 2022; WWC, 2022). Many scholarly journals published special issues to devote exclusively to these advanced approaches, such as *Journal of School Psychology* (2014, Volume 52, Issue 2), *Remedial and Special Education* (2017, Volume 38, Issue 6), *Developmental Neurorehabilitation* (2018, Volume 21, Issue 4), and *Perspectives on Behavior Science* (2022, Volume 45, Issue 1). The advanced approaches include (1) quantifying an intervention effect with standardized/unstandardized indices (e.g., Hedges et al., 2012, 2013; Moeyaert et al., 2013; Pustejovsky et al., 2014; Ugille et al., 2012, 2014) or with non-overlapping ESs (e.g., Michiels & Onghena, 2019; Parker & Vannest, 2009), (2) employing different methods to estimate an intervention effect (e.g., the method of moments by Hedges et al., 2012, 2013; the restricted maximum likelihood method by Pustejovsky et al., 2014; the Bayesian method by Natesan, 2019 or Natesan & Hedges, 2017), and (3) inferring an intervention effect based on a statistical model (e.g., hierarchical linear modeling by Pustejovsky et al., 2014) or a design (e.g., randomization tests by Michiels & Onghena, 2019, and Onghena, 2020). Among the advanced approaches, design-comparable ESs (D-CESs) were proposed as standardized indices to synthesize intervention effects across SCED and group studies, or over different outcome measures, based on a statistical model (Hedges et al., 2012, 2013; Pustejovsky et al., 2014; Shadish et al., 2014; Zelinsky & Shadish, 2018).

Version 5.0 of the *What Works Clearinghouse Procedures and Standards Handbook* (WWC, 2022) specifically recommends reporting D-CES indices, along with visual analysis, when assessing an intervention effect in SCED studies. Yet to the best of our knowledge, no published research has investigated D-CES's statistical assumptions in SCED contexts. If statistical assumptions, such as normality, are not met or not robust, inferences derived from D-CES lack statistical validity. Furthermore, the small sample sizes and limited number of measurements used in most SCED studies render the normality assumption unlikely to be robust, if it is violated.

Small sample sizes and limited number of measurements are also a central concern when an effective intervention is to be generalized to another sample, setting, location, behavior, or measurement (Horner et al., 2005). That is, how does one know that an intervention is generalizable beyond the effect already documented in a SCED study? One way to address, or even enhance, the generalizability of an effect is to systematically replicate an intervention in different contexts using different participants, behaviors, and parallel measurements (Horner et al., 2005; Kazdin, 2011). Replication results are subsequently synthesized using meta-analysis methods (Becraft et al., 2020; Beretvas & Chung, 2008; Moeyaert et al., 2020; Onghena et al., 2018). To this end, methodologists have devised various approaches to perform SCED meta-analysis (Jamshidi et al., 2022; Vlaeyen et al., 2020). According to Becraft et al. (2020) and Moeyaert et al. (2021), there has been a dramatic increase from 1987 to 2019 in the number of scholarly publications on SCED intervention studies and their meta-analyses.

Definition of g_{AB} for SCED studies

One D-CES[1], namely g_{AB}, was proposed by Pustejovsky et al. (2014) to quantify intervention effects within and across primary SCED studies (Hedges et al., 2012, 2013; Pustejovsky et al., 2014; Shadish et al., 2014; WWC, 2022; Zelinsky & Shadish, 2018). A cursory search of the literature since 2014 has found g_{AB} reported in numerous primary studies or meta-analysis (Anaby et al., 2020; Grasley-Boy et al., 2021; Lee et al., 2022; Peltier et al., 2021; Peltier et al., 2020a, b; Rincón et al., 2021; Rivera Pérez et al., 2022; Romano & Windsor, 2020; Romano et al., 2021; Ruiz et al., 2018; Saul & Norbury, 2021; Teh et al., 2021; Thurmann-Moe et al., 2021). g_{AB} is the D-CES specifically recommended by the *What Works Clearinghouse Procedures and Standards Handbook, Version 5.0* (WWC, 2022) for SCED studies.

g_{AB} is a sample estimator of the population standardized mean difference (δ_{AB}) between an A phase and a B phase, similar to Cohen's *d* or Hedges' *g* used in group studies (Pustejovsky et al., 2014; WWC, 2022). The δ_{AB} is defined according to Eq. 1:

$$\delta_{AB} = \frac{\mu_B - \mu_A}{\sqrt{\sigma^2 + \tau^2}}, \quad (1)$$

where μ_A is the population mean of Phase A measurements, μ_B is the population mean of Phase B measurements, σ^2 is the variance of measurements within cases, and τ^2 is the variance of measurements across cases. Thus, $(\sigma^2 + \tau^2)$ is the total variance of measurements within and across all cases.

[1] A D-CES is also called between-case standardized mean difference (BC-SMD) or simply standardized mean difference (SMD) in the literature (Barton et al., 2019; Moeyaert et al., 2021; Peltier et al., 2020b; Valentine et al., 2016).

Among the three MB design variations mentioned earlier, g_{AB} is suitable only for MB designs across three or more cases of the same behavior. It is a product of a bias correction factor [$J(\nu)$] and the sample estimate of δ_{AB} ($\hat{\delta}_{AB}$), as in Eq. 2:

$$g_{AB} = J(\nu) \times \hat{\delta}_{AB}. \quad (2)$$

Both $J(\nu)$ and $\hat{\delta}_{AB}$ are estimated by the restricted maximum likelihood (REML) method (Pustejovsky et al., 2014) which we explain in the "Method" section. The performance of g_{AB} under normal and non-normal distributions is the focus of the present simulation study.

Five MB models formulated by Pustejovsky et al. (2014)

Pustejovsky et al. (2014) formulated five models for MB data. They are sequentially named MB1 to MB5 in this paper. All five models permit cases to vary in Phase A levels. They differ in how the intervention effect and the trends in A or B phase are modeled across cases. The five MB models are hierarchical linear models in which Level-1 parameters model the individual data and Level-2 parameters model how the Level-1 parameters vary across cases. Being the simplest and most restrictive model, MB1 assumes a fixed intervention effect for all cases with no trend in either A or B phase. MB1 is recommended by the *What Works Clearinghouse Procedures and Standards Handbook, Version 5.0* (WWC, 2022) as a "starting point" (p. 182) for assessing an intervention effect.

MB2 to MB5 are more flexible and extensible than MB1, due to additional parameters and fewer restrictions (Pustejovsky et al., 2014). MB2 assumes a varying immediate effect due to intervention across cases, with no trend in either A or B phase. MB3 assumes a fixed intervention effect with a fixed linear trend in either A, B, or both phases. MB4 assumes a fixed intervention effect with a varying linear trend in the A phase and a fixed linear trend in the B phase. Being the most complex model, MB5 assumes a fixed intervention effect with a varying linear trend in both A and B phases. It is worth noting that MB2 is the only model among the five proposed that allows the immediate effect of intervention to vary across cases. Pustejovsky et al. (2014) conducted three simulation studies under MB1, MB2, and MB4 to provide empirical evidence to support the reporting of g_{AB}. Results from the three simulation studies are summarized next.

Three simulation studies of g_{AB} under MB1, MB2, and MB4

The first simulation study (Study 1) was conducted under MB1, Study 2 under MB2, and Study 3 under MB4. In all three studies, data were simulated from normal distributions. For Studies 1 and 2, four levels of number of cases (m) were used: 3, 4, 5, and 6. For Study 3, two additional levels were added to m: 3, 4, 5, 6, 9, and 12. The number of measurements (N) was either 8 or 16, the within-case reliability (ρ = the ratio of between-case variance to the total variance within and between cases) ranged from 0 to 0.8, and the first-order autocorrelation (ϕ) ranged from −0.7 to 0.7 in Studies 1 to 3. For Studies 2 and 3, the ratio of variance components (λ) was either 0.1 or 0.5. λ was defined in Study 2 as the variance of all cases' level shifts between A and B phases as a fraction of the variance of all cases' Phase A levels. In Study 3, λ was defined as the variance of all cases' baseline slopes as a fraction of the variance of all cases' Phase A levels. Four criteria were used in all three studies to assess the performance of g_{AB}: relative bias, relative bias of variance estimators, *MSE*, and coverage rate of the 95% CIs. The 95% CI was constructed using two methods: the symmetric and the noncentral t.

Results from Study 1 of Pustejovsky et al. (2014) showed that relative bias of g_{AB} under MB1 was small. At the smallest $m = 3$ and $N = 8$, the relative bias was no more than 4.3%, yet the relative bias of g_{AB}'s variance estimator was 16%. As both m and N increased, g_{AB}'s variance estimate was very close to the true variance. Between the two CI methods, the average coverage rate of the symmetric method was closer to the nominal level of 95% than the noncentral t method.

Results from Study 2 showed that g_{AB}'s average relative bias was small under MB2. Relative bias was generally greater when $N = 8$ than when $N = 16$. At the smallest $m = 3$ and $N = 8$, the relative bias was no more than 7.3%. For $m = 4$, the relative bias was always less than 4.9%. The relative bias decreased to no more than 2.9% when $m \geq 5$. The variance of g_{AB} was overestimated. The relative bias in g_{AB}'s variance estimator was as large as 43% when $m = 3$ and $N = 16$. Even when $m = 6$ (the largest under MB2) and $N = 16$, the relative bias was still 14%. The *MSE*s under MB2 ranged from 0.092 when $m = 6$ and $N = 16$, to 0.290 when $m = 3$ and $N = 8$. The *MSE*s generally increased as ρ, λ, and ϕ increased. Between the two CI methods, the symmetric method maintained an average coverage rate closer to 95% than the noncentral t method. Based on these results, Pustejovsky et al. (2014) recommended g_{AB} for meta-analysis with $m \geq 4$ and the symmetric method for constructing CIs of g_{AB} under MB2.

Results from Study 3 revealed the same pattern under MB4 as under MB2, namely small relative bias. As with results of MB2, g_{AB} as a point estimator was suitable for studies with $m \geq 4$. *MSE* obtained under MB4 was large, compared with those obtained under MB1, especially when m was small. Unlike results obtained under MB1 and MB2, g_{AB}'s variance was underestimated, except when $m = 3$ and $N = 8$. The variance's underestimation was more pronounced when $N = 16$ than when $N = 8$. The average *MSE* of g_{AB} under MB4 ranged from 0.066 when $m = 12$ and $N = 16$, to 0.596 when $m = 3$ and $N = 8$. For a given m and N, *MSE* derived under MB4 were

larger than those under MB1 or MB2, especially when m was small. *MSE*s generally increased as ρ, λ, and ϕ increased. The 95% CI based on the symmetric method approached the nominal level when m was large (i.e., ≥ 9). The CI based on the noncentral t method tended to substantially undercover the population δ_{AB} when $m = 3, 4, 5, 6$, or 9.

Based on Studies 1 to 3, Pustejovsky et al. (2014) concluded that the relative bias of g_{AB} was reasonably small, even with very few cases. Yet large sample sizes were needed in order to yield precise point estimates, reasonably accurate *SE* estimates and CIs under a complex model, namely MB4. Pustejovsky et al. (2014) further cautioned not to rely on model-based *SE* estimates in meta-analysis, because inaccurate *SE* estimates lead to inaccurate weights for primary studies and inaccurate estimates of between-study heterogeneity for meta-analysis.

The normality assumption of g_{AB} and the REML method

As previously mentioned, data in Studies 1 to 3 of Pustejovsky et al. (2014) were simulated only from normal distributions. Indeed, g_{AB} assumes that data within and across cases are normally distributed. Yet non-normal data are quite common in SCED studies (e.g., Au et al., 2017; Brosnan et al., 2018; Ferron et al., 2014; Joo, 2017; Stewart & Hall, 2017). Furthermore, due to asymptotic normality[2] of the REML method, voluminous data are needed in order to yield an acceptable g_{AB} for δ_{AB}. The ML method, of which the REML is a special case, is known to perform poorly when data are limited, even under normal conditions (Braunstein, 1992). Yet small sample sizes (or cases) and limited numbers of measurements are the norm rather than an exception in SCED studies. According to Shadish and Sullivan (2011), 73.5% of 809 SCED studies employed one to 13 participants with an average of 3.64 cases per study. Tanious and Onghena (2021) reported that the median sample size used in 210 MB studies was 4 with an interquartile range of 4. As for the number of measurements, Shadish and Sullivan (2011) reported that 90.6% of 809 SCED studies used 49 or fewer measurements with a median of 20 measurements. Pustejovsky et al.'s (2019) review of 303 SCED studies found a median of 7 measurements in initial baseline phases, with an interquartile range of 7.

Pustejovsky et al. (2014) did not investigate the performance of g_{AB} under non-normal conditions. Furthermore, sample sizes and the number of measurements used in their simulation studies were small for the REML method. It remains unknown whether g_{AB} performs satisfactorily under non-normal conditions with small samples and limited numbers of measurements (Maas & Hox, 2004; Man et al., 2022; Raudenbush & Bryk, 2002). To the best of our knowledge, no published study has systematically investigated the singular impact of non-normality, or the joint impact of non-normality with other data features (e.g., sample size, autocorrelation), on the performance of g_{AB} in primary SCED studies and their meta-analyses.

[2] Asymptotic normality is a property of the REML or ML method in which the normal approximation to the sampling distribution of a ML estimator (or g_{AB} in this study) is valid when an asymptotically large sample of data is used (McNeish, 2017).

Aims of the present simulation

The present study aimed to fill the voids in the literature by investigating how distributions of data singularly and jointly impacted the performance of g_{AB} under MB2. We focused on MB2 because MB2 is more flexible, but less researched, than MB1. MB2 is also the only model among the five proposed by Pustejovsky et al. (2014) that allows the immediate effect of an intervention to vary across cases.

The singular and joint impacts of data distribution on g_{AB} were investigated by simulating data from normal and non-normal distributions, and by manipulating five data features (number of cases, number of measurements, autocorrelation, within-case reliability, and ratio of variance components). The five data features were also manipulated in Pustejovsky et al. (2014). The performance of g_{AB} was evaluated by the same four criteria as Pustejovsky et al. (2014): relative bias, relative bias of variance, *MSE*, and coverage rate. These four criteria have been routinely used to assess a statistic (e.g., g_{AB}) in primary studies (e.g., Algina et al., 2005; Hoogland & Boomsma, 1998), or for meta-analysis (e.g., American Psychological Association, 2020; Hoogland & Boomsma, 1998; Pustejovsky et al., 2014). Based on the evaluation of the four criteria, we identified conditions in which g_{AB} performed acceptably for primary MB studies and meta-analysis.

In sum, the present study aimed to answer two research questions under MB2:

RQ1: What is the impact of data distribution, number of cases, number of measurements, within-case reliability, ratio of variance component, and autocorrelation on the performance of g_{AB} as measured by relative bias, relative bias of variance, *MSE*, and coverage rate?
RQ2: What are the conditions in which g_{AB} performed acceptably for primary MB studies and meta-analysis?

Findings from the present study should provide empirical evidence to extend the recommendation made by the *What Works Clearinghouse Procedures and Standards Handbook, Version 5.0* (WWC, 2022) to MB2. They should also inform practitioners and researchers about the suitability of g_{AB} for MB studies and their meta-analysis. To this end, we provide general recommendations on conditions under which it is appropriate to use g_{AB} to assess intervention effects. This paper concludes with a discussion of g_{AB}'s applicability in SCED contexts, its design-comparability across SCED and group studies, and sound reporting practices of ES indices including g_{AB}.

Method

In this section, we present MB2 and its assumptions first, followed by the definition of the population standardized mean difference (δ_{AB}) under MB2. Next, we describe the simulation design, justifications for the manipulated conditions, and an outline of seven steps for simulating and analyzing simulated data. Details of each step are provided following the outline.

MB2

As previously stated, MB2 assumes that cases vary in the average score of the A phase and also in the immediate intervention effect between A and B phases. Pustejovsky et al. (2014) referred to the average score of the A phase as Phase A level, and the immediate intervention effect as a level shift between A and B phases. MB2 also assumes that there is no linear trend in either A or B phase (Pustejovsky et al., 2014). Thus, for the *j*th measurement of the *i*th case, the score Y_{ij} is modeled by a within-case model according to Eq. 3:

$$Y_{ij} = \beta_{0i} + \beta_{1i} \times D_{ij} + \varepsilon_{ij}, \tag{3}$$

where β_{0i} = Phase A level for Case *i*; β_{1i} = level shift for Case *i* = the immediate change in Case *i*'s measurement due to intervention; D_{ij} = a dummy variable that equals 0 (for Phase A measurements) or 1 (for Phase B measurements); ε_{ij} = Level-1 error; $i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, N$; m = the number of cases; and N = the total number of measurements in A and B phases combined[3].

Because MB2 assumes a varying Phase A level and a varying level shift between A and B phases across cases, β_{0i} and β_{1i} are further modeled by a between-case model according to Eqs. 4 and 5:

$$\beta_{0i} = \gamma_{00} + \eta_{0i}, \tag{4}$$

$$\beta_{1i} = \gamma_{10} + \eta_{1i}, \tag{5}$$

where γ_{00} = the average Phase A level, γ_{10} = the average level shift between A and B phases, and η_{0i} and η_{1i} are Level-2 errors.

Substituting Eq. 4 for β_{0i} and Eq. 5 for β_{1i} into Eq. 3, we obtain Eq. 6 of fixed and random effects for the distribution of Y_{ij}—the *j*th measurement of the *i*th case—under MB2:

$$\begin{aligned} Y_{ij} &= \gamma_{00} + \eta_{0i} + \left(\gamma_{10} + \eta_{1i}\right) \times D_{ij} + \varepsilon_{ij} \\ &= \left[\gamma_{00} + \left(\gamma_{10} \times D_{ij}\right)\right] + \left\{\eta_{0i} + \left(\eta_{1i} \times D_{ij}\right) + \varepsilon_{ij}\right\} \\ &= [\text{fixed effects}] + \{\text{random effects}\}. \end{aligned} \tag{6}$$

[3] *N* refers to the total number of measurements per case. It was set to be identical for all cases in each condition.

Statistical and design assumptions for MB2 are stated in (a) to (f) below, according to Pustejovsky et al. (2014). The present study investigated normality assumptions stated in (a) and (d).

(a) Within cases, ε_{ij}s are normally distributed with a mean of 0 and a variance of σ^2.
(b) Within cases, ε_{ij}s are correlated with a first-order autocorrelation ϕ, or Cov (ε_{ij}, ε_{ik}) = $\phi^{|k-j|}\sigma^2$.
(c) Across cases, ε_{ij}s are homoscedastic and independently distributed, namely, $\text{Var}(\varepsilon_{ij}) = \text{Var}(\varepsilon_{hk}) = \sigma^2$ and Cov (ε_{ij}, ε_{hk}) = 0 for all $i \neq h$.
(d) (η_{0i}, η_{1i}) are multivariate normally distributed with mean (0, 0) and a covariance matrix $\mathbf{T_{2\times2}} = \begin{bmatrix} \tau_0^2 & \tau_{10} \\ \tau_{10} & \tau_1^2 \end{bmatrix}$, where τ_0^2 is the variance of all cases' Phase A levels, τ_1^2 is the variance of all cases' level shifts between A and B phases, and τ_{10} is the covariance between Phase A levels and level shifts[4].
(e) Level-1 errors (ε_{ij}s) are independent of Level-2 errors (η_{0i} and η_{1i}).
(f) Measurements are equally spaced over time.

Definition of δ_{AB}

Under MB2, the population mean difference = $(\gamma_{00} + \gamma_{10}) - \gamma_{00} = \gamma_{10}$, and the total variance within and across cases = $\sigma^2 + \tau_0^2$. Hence, δ_{AB} is defined by Eq. 7:

$$\delta_{AB} = \frac{\gamma_{10}}{\sqrt{\sigma^2 + \tau_0^2}}. \tag{7}$$

Equation 7 is identical to Eq. 1, except for the notation differences (Pustejovsky et al., 2014). Guided by Pustejovsky et al. (2014), we set[5] $\gamma_{00} = 0$, $\gamma_{10} = 1$, and $\sigma^2 + \tau_0^2 = 1$. Therefore, $\delta_{AB} = 1$ for all simulated conditions in this study and also in Studies 1 to 3 of Pustejovsky et al. (2014).

Simulation design

The present study manipulated six factors according to Table 1. The first factor (Dist or distribution of data) was unique to the present study. The next four factors, namely, *m*, *N*, ρ, and λ were manipulated identically[6] as in Study 2 of Pustejovsky et al. (2014). The sixth factor (ϕ) was manipulated slightly differently from Study 2 of Pustejovsky et al. (2014). Justifications for manipulated conditions are given in

[4] The covariance between Phase A levels and level shifts (τ_{10}) was set to 0 in this study, as Pustejovsky et al. (2014) did.

[5] For the purposes of this simulation study, we set $\sigma^2 + \tau_0^2 = 1$, as Pustejovsky et al. (2014) did. σ^2 and τ_0^2 were not specified separately.

[6] The present study did not include $\rho = 0$, because $\rho = 0$ leads to an undefined λ. Study 2 of Pustejovsky et al. (2014) included $\rho = 0$.

Table 1 Simulation design of the present study

Factor	Definition	No. of levels	Conditions
Dist	Distribution of data (skewness, kurtosis[a])	4	normal (0, 0) nearly normal (0, 0.35) mildly non-normal (1, 0.35) moderately non-normal (1, 3)
m	Number of cases	4	3, 4, 5, 6
N	Number of measurements	2	8, 16
ρ	Within-case reliability $= \tau_0^2/(\sigma^2 + \tau_0^2)$	4	0.2, 0.4, 0.6, 0.8
λ	Ratio of variance components $= \tau_1^2/\tau_0^2$	2	0.1, 0.5
ϕ	First-order autocorrelation	7	−0.4, −0.3, −0.1, 0, 0.1, 0.3, 0.4

Note. [a]Kurtiosis is defined in this paper as the fourth moment of a distribution minus 3. This kurtosis is also called excess kurtosis (Christoffersen, 2004; Cotter & Hanly, 2012; Darbyshire & Hampton, 2012)

the next section. A total of 1792 conditions (= 4 × 4 × 2 × 4 × 2 × 7) were manipulated. Table 2 presents the start points for the intervention across cases. The start points were identical to those used in Study 2 of Pustejovsky et al. (2014).

Justifications for manipulated conditions

The distribution of data was manipulated through the joint manipulation of Level-1 and Level-2 error distributions in Eq. 6. Four distributions—one normal and three non-normal—were simulated as the distributions of sums of Level-1 and Level-2 errors. Because of the large number of conditions (= 1792) investigated in this study, we did not simulate Level-1 and Level-2 errors separately from two different distributions (e.g., normal for Level-1 errors and non-normal for Level-2 errors). Each distribution was specified through the specification of its skewness and kurtosis (Joo & Ferron, 2019; Man et al., 2022; Owens & Farmer, 2013). For the normal distribution, we specified skewness = kurtosis = 0. For the nearly normal distribution, skewness = 0 and kurtosis = 0.35 were specified. For the mildly non-normal distribution, skewness = 1 and kurtosis = 0.35 were specified. For the moderately non-normal distribution, we specified skewness = 1 and kurtosis = 3. The four marginal distributions are shown in File 1[7] at https://osf.io/hsvwu/.

Table 2 Start points for intervention in N measurements

m	Case number	Start point in $N = 8$	Start point in $N = 16$
3	Case 1	4	5
	Case 2	5	9
	Case 3	6	13
4	Case 1	4	4
	Case 2	4	7
	Case 3	5	11
	Case 4	6	14
5	Case 1	4	4
	Case 2	4	6
	Case 3	5	9
	Case 4	5	12
	Case 5	6	14
6	Case 1	4	4
	Case 2	4	6
	Case 3	5	8
	Case 4	5	10
	Case 5	6	12
	Case 6	6	14

We decided on these four distributions on the basis of empirical skewness and kurtosis of SCED data (Joo, 2017; Solomon, 2014) and conditions manipulated in Owens and Farmer (2013). Joo (2017) reported empirical skewness to range from −0.71 to 1.91 and empirical kurtosis from −1.07 to 3.01, based on 20 MB data sets published in the *Journal of Applied Behavior Analysis*. Solomon (2014) reported empirical skewness to range from 0.46 to 2.89 and empirical kurtosis from 0.49 to 1.57, based on 104 SCED studies of school-based interventions. Owens and Farmer (2013) investigated Level-2 normality assumption

[7] Supplemental materials of the present study include nine files; they are available at https://osf.io/hsvwu/. File 1 presents the four marginal distributions. File 2 summarizes procedures used to confirm data generation under non-normal distributions. File 3 explains the decisions for the seven levels of ϕ. File 4 presents the modified R scdhlm package. File 4.1 is the superordinate R program to establish simulation conditions, execute File 4, and check convergence of each simulation. File 5 contains the actual values of the four criteria under each condition. File 6 presents alternative three-way ANOVA results including interactions of Dist with two of the other five factors in the model. File 7 presents the P_{25}, P_{50}, mean, P_{75}, P_{95} of *MSE*s under each combination of m and ρ. File 8 presents results from converged and non-converged replications and non-convergence rates.

for MB data using multilevel modeling. They manipulated six Level-2 unimodal distributions ranging from normal (skewness = 0, kurtosis = 0), (0, −1), (0, 2), (0, 3.75), (1, 2), to most non-normal (1, 3.75). Of the six, four were symmetric and two were positively skewed. Of the four unimodal distributions manipulated in this study, two were symmetric and two were positively skewed. The three non-normal distributions of the present study were specified with skewness and kurtosis well within their respective empirical ranges reported in Joo (2017) and Solomon (2014). We confirmed that skewness and kurtosis of our simulated data matched closely with those specified in Table 1 (see File 2 at https://osf.io/hsvwu/).

Regarding *m* and *N*, Pustejovsky et al. (2014, supplemental materials[8]) justified their conditions by spreading the intervention start points as evenly as possible over measurements, while keeping at least three measurements in each phase. The within-case reliability ($\rho = \tau_0^2/\left(\sigma^2 + \tau_0^2\right)$, also called the intra-class correlation or ICC) was varied from 0.2 to 0.8 in increments of 0.2. A ρ of 0.2 represented a low between-case variance in levels (τ_0^2) relative to the within-case variance (σ^2), hence a low within-case reliability. A ρ of 0.8 represented a high between-case variance in levels relative to the within-case variance, hence a high within-case reliability. The ratio of variance components ($\lambda = \tau_1^2/\tau_0^2$) was set to either 0.1 or 0.5. According to Pustejovsky et al. (2014, supplementary materials), a λ of 0.1 represented a moderate level of the between-case variation in level shifts, relative to the between-case variation in Phase A levels. A λ of 0.5 represented a high level of the between-case variation in level shifts, relative to the between-case variation in Phase A levels.

Because of the repeated observation of the same behavior in SCED studies, each case's measurements are correlated. Such a correlation is quantified by the first-order autocorrelation (ϕ). Pustejovsky et al. (2014) manipulated the autocorrelation under MB2 to range from −0.7 to 0.7, in increments of 0.2. To ensure that our manipulation of autocorrelation was plausible for MB2 under non-normal distributions, we tested a range of autocorrelations based on empirical and simulation studies (Joo, 2017; Joo & Ferron, 2019; Solomon, 2014). We eventually decided on seven levels of ϕ: −0.4, −0.3, −0.1, 0, 0.1, 0.3, 0.4 for the present study (see File 3).

For each of the 1792 conditions, 20,000 replications were generated. We modified the R scdhlm package (Pustejovsky et al., 2021) for this simulation study. The modified R scdhlm package can be found in File 4 at https://osf.io/hsvwu/. File 4.1 is the superordinate R program to establish simulation conditions, execute File 4, and check convergence of each simulation. Under normal distributions, the modified R scdhlm package produced results comparable to those obtained from Study 2 of Pustejovsky et al. (2014) (see Appendix A).

The simulation and analysis procedures are outlined in seven steps. Details on each step are presented following the outline.

[8] Pustejovsky et al.'s (2014) supplemental materials are available from https://www.jepusto.com/files/Effect-sizes-in-multiple-baseline-designs-Simulation-results.pdf.

Outline of simulation and analysis procedures

Step 1: Generate 20,000 random seeds which were used to create 20,000 replications for the 1792 conditions.
Step 2: Given a random seed from Step 1, simulate a replication under a specific condition of MB2.
Step 3: Use the REML method to compute g_{AB} based on data generated in Step 2.
Step 4: Repeat Steps 2 and 3 until 20,000 replications and 20,000 g_{AB}s were obtained for each of the 1792 conditions.
Step 5: Compute four criteria as indicators of the performance of g_{AB}.
Step 6: Analyze the impact of the six factors on the four criteria.
Step 7: Identify conditions in which g_{AB} performed acceptably for MB studies and meta-analysis.

Step 1: Generate 20,000 random seeds for the 1792 conditions

Before the simulation began, we assessed the adequacy of $R = 20{,}000$ replications used in Pustejovsky et al. (2014) by examining its Monte Carlo *SE*. A Monte Carlo *SE* provides an estimate of the empirical *SE* resulted from *R* replications. The Monte Carlo *SE* for the expected coverage rate of 95% CIs based on 20,000 replications was computed according to Eq. 8:

$$\text{Monte Carlo } SE = \left[0.95 \times (1 - 0.95)/20{,}000\right]^{0.5} = 0.00154, \quad (8)$$

where 0.95 = the expected coverage rate of 95% CIs. Such a Monte Carlo *SE* was deemed acceptable by Morris et al. (2019), who suggested keeping the Monte Carlo *SE* below 0.005. We therefore considered 20,000 replications adequate for the present study. And 20,000 random seeds were generated and used in each condition.

Step 2: Simulate a replication under a specific condition

Given Eq. 6 as the distribution of Y_{ij} and start points in Table 2, we generated data from one of the four distributions specified in Table 1. For *N* scores of a case, Eq. 6 can be expressed in matrix notations as Eq. 9:

$$\mathbf{Y}_{N\times1} = \mathbf{D}_{N\times2}\,\boldsymbol{\gamma}_{2\times1} + \mathbf{e}_{N\times1}. \quad (9)$$

The fixed effects in Eq. 6 are expressed as the product of the design matrix ($\mathbf{D}_{N\times2}$) and a fixed-effect vector ($\boldsymbol{\gamma}_{2\times1}$) in Eq. 9. The random effects in Eq. 6 are expressed as an error vector ($\mathbf{e}_{N\times1}$) in Eq. 9. The $\mathbf{e}_{N\times1}$ vector consists of Level-2 errors [$\eta_{0i} + (\eta_{1i} \times D_{ij})$ in Eq. 6] and Level-1 errors (ε_{ij} in Eqs. 3 and 6).

As previously stated, sums of Level-1 and Level-2 errors followed a normal or non-normal distribution that was specified by its skewness and kurtosis. To generate random errors of $\mathbf{e}_{N\times1}$ from a multivariate normal distribution, we specified skewness = 0, kurtosis = 0, and a variance-covariance matrix of errors ($\boldsymbol{\Sigma}_{N\times N}$) in the mvrnonnorm function of the semTools package (Jorgensen et al., 2021). The $\boldsymbol{\Sigma}_{N\times N}$ is written in matrix notation as Eq. 10:

$$\boldsymbol{\Sigma}_{N\times N} = \mathbf{DTD^T} + (1-\rho) \bullet \mathbf{AR(1)}, \quad (10)$$

where $\mathbf{DTD^T}$ = the variance-covariance matrix of Level-2 errors, $(1-\rho)\bullet\mathbf{AR(1)}$ = the variance-covariance matrix of Level-1 errors, $\mathbf{D}$ = design matrix from Eq. 9, $\mathbf{T}_{2\times2} = \begin{bmatrix} \rho & 0 \\ 0 & \rho\times\lambda \end{bmatrix} = \begin{bmatrix} \tau_0^2 & 0 \\ 0 & \tau_1^2 \end{bmatrix}$ (see Footnote 4), and $\mathbf{AR(1)}$ is the matrix of first-order autocorrelations with 1s along the diagonal and $\phi^{|k-jl}$ off-diagonal. Once skewness, kurtosis, and $\boldsymbol{\Sigma}_{N\times N}$ were specified, the mvrnonnorm function produced multivariate normal errors using the Vale and Maurelli method (Vale & Maurelli, 1983). Appendix B describes details in generating a replication of 3 (=m) cases from a normal distribution (Dist = normal) with 8 (=N) measurements, within-case reliability (ρ) = 0.2, ratio of variance components (λ) = 0.1, and first-order autocorrelation (ϕ) = −0.4.

Errors were similarly generated from the other three distributions by specifying their corresponding skewness and kurtosis, plus a $\boldsymbol{\Sigma}_{N\times N}$ in the mvrnonnorm function (see File 4 at https://osf.io/hsvwu/). After data were generated for m cases, a replication was formed and a g_{AB} was computed.

Step 3: Use the REML method to compute g_{AB}

To explain the details of Step 3, we reformulate δ_{AB} in matrix notation. Next, we describe the estimation of δ_{AB} by g_{AB} and the estimation of the variance of g_{AB}.

Reformulating δ_{AB} in matrix notation

Using Pustejovsky et al.'s (2014) matrix notations, we define the vector of fixed effects of MB2 as $\boldsymbol{\gamma}_{2\times1} = (\gamma_{00}, \gamma_{10})^T$ and the vector of variance components as $\boldsymbol{\omega}_{5\times1} = (\sigma^2, \phi, \tau_0^2, \tau_1^2, \tau_{10})^T$. The $\boldsymbol{\omega}_{5\times1}$ vector includes the within-case variance σ^2, the first-order autocorrelation ϕ, Level-2 variances τ_0^2 and τ_1^2, and the covariance τ_{10} (see Footnote 4). With two constant vectors defined as $\mathbf{p}_{2\times1} = (0, 1)^T$ and $\mathbf{r}_{5\times1} = (1, 0, 1, 0, 0)^T$, the δ_{AB} of Eq. 7 is reformulated as Eq. 11:

$$\delta_{AB} = \frac{\mathbf{p}^T\boldsymbol{\gamma}}{\sqrt{\mathbf{r}^T\boldsymbol{\omega}}}. \quad (11)$$

Estimating δ_{AB} by g_{AB}

g_{AB} is the product of the bias correction factor, $J(\nu)$, multiplied with the REML estimate of δ_{AB} (i.e., $\hat{\delta}_{AB}$), as in Eq. 12:

$$g_{AB} = J(\nu) \times \hat{\delta}_{AB}, \quad (12)$$

where $J(\nu) = 1 - 3/(4\nu - 1)$ and ν is determined from Eq. 13.

$$\nu = \frac{2\left(\mathbf{r}^T\hat{\boldsymbol{\omega}}\right)^2}{\mathbf{r}^T\mathbf{C}(\hat{\boldsymbol{\omega}})\mathbf{r}}, \quad (13)$$

where $\mathbf{C}(\hat{\boldsymbol{\omega}})$ is the estimated covariance matrix of $\hat{\boldsymbol{\omega}}$, and $\hat{\boldsymbol{\omega}}$ is the REML estimate of $\boldsymbol{\omega}$. When m and N both approach infinity, $\hat{\boldsymbol{\omega}}$ approaches $\boldsymbol{\omega}$, $\mathbf{C}(\hat{\boldsymbol{\omega}})$ approaches a null matrix, ν approaches infinity, and $J(\nu)$ approaches 1; hence, the need for bias correction diminishes.

By plugging $\boldsymbol{\gamma}$'s REML estimate ($\hat{\boldsymbol{\gamma}}$) and $\hat{\boldsymbol{\omega}}$ into Eq. 11, we obtain $\hat{\delta}_{AB}$ from Eq. 14:

$$\hat{\delta}_{AB} = \frac{\mathbf{p}^T\hat{\boldsymbol{\gamma}}}{\sqrt{\mathbf{r}^T\hat{\boldsymbol{\omega}}}}. \quad (14)$$

The REML algorithm estimated the random effects (i.e., $\boldsymbol{\omega}$) iteratively using a non-linear maximization approach. The algorithm stopped when it met a pre-specified convergence criterion (i.e., tolerance = 10^{-6}), or when it reached a pre-specified number of iterations (=50). If the REML algorithm did not converge to the convergence criterion after 50 iterations, we re-simulated data (see File 4.1)[9]. After obtaining

[9] Our decision to discard non-converged results was guided by Paxton et al. (2001) who stated, "If the purpose of the Monte Carlo analysis is to provide realistic information to users of the technique, then non-converged samples, which are rarely assessed in practice, will provide irrelevant information and subsequently threaten external validity" (pp. 301–302). In other words, a non-converged result could not be viewed as an optimal REML estimate of δ_{AB}. Authors of several simulation studies held similar views on non-converged results, including Bandalos and Leite (2013, p. 655), Bolin et al. (2019, p. 226), Bollen et al. (2014, p. 7), and Fan and Fan (2005, p. 131). Hence, we decided to retain only converged results until we obtained 20,000 replications in each condition. File 8 presents results from converged and non-converged replications and non-convergence rates.

the estimated random effects ($\hat{\boldsymbol{\omega}}$), the algorithm estimated fixed effects ($\hat{\boldsymbol{\gamma}}$) using the generalized least squares estimator (Jiang, 2007).

Estimating the variance of g_{AB}

The variance of g_{AB} is estimated from Eq. 15 (Hedges, 2007; Pustejovsky et al., 2014):

$$V_{g_{AB}} = J(\nu)^2\left[\frac{\nu\kappa^2}{\nu-2} + g_{AB}^2 \times \left(\frac{\nu}{\nu-2} - \frac{1}{J(\nu)^2}\right)\right], \tag{15}$$

where $J(\nu)$ and g_{AB} are computed from Eqs. 12 and 14, ν by Eq. 13, and κ by Eq. 16.

$$\kappa = \sqrt{\frac{\mathbf{p}^T\mathbf{C}(\hat{\boldsymbol{\gamma}})\mathbf{p}}{\mathbf{r}^T\hat{\boldsymbol{\omega}}}}, \tag{16}$$

where $\mathbf{C}(\hat{\boldsymbol{\gamma}})$ is the estimated covariance matrix of $\hat{\boldsymbol{\gamma}}$.

Step 4: Repeat Steps 2 and 3 until 20,000 replications and 20,000 g_{AB}s are obtained

Steps 2 and 3 were repeated until all the data for the present study were generated. At the end of this step, we obtained 20,000 replications and 20,000 g_{AB}s in each of the 1792 conditions.

Step 5: Compute four criteria

We applied the same four criteria as those used in Pustejovsky et al. (2014) to assess the performance of g_{AB}. The four criteria were relative bias, relative bias of g_{AB}'s variance estimator, *MSE*, and coverage rate of symmetric 95% CI. They are abbreviated as *RB*, *RBV*, *MSE*, and *CR* respectively. Each criterion is defined below.

RB (relative bias)

The *RB* of g_{AB} was calculated according to Eq. 17:

$$RB = \frac{\overline{g}_{AB} - \delta_{AB}}{\delta_{AB}}, \tag{17}$$

where $\overline{g}_{AB}$ was the mean of 20,000 g_{AB}s obtained in each condition. Because $\delta_{AB} = 1$, bias and *RB* were the same. We refer to them both as *RB*. Based on Hoogland and Boomsma (1998), we interpreted $|RB| < 5\%$ as acceptable and $|RB| \geq 5\%$ as unacceptable. In addition, $RB < -5\%$ was interpreted as unacceptable underestimate and $RB > 5\%$ as unacceptable overestimate.

RBV (relative bias of g_{AB}'s variance estimator)

The *RBV* of g_{AB} was calculated according to Eq. 18[10],

$$RBV = \frac{\overline{V}_{g_{AB}} - \mathrm{Var}(g_{AB})}{\mathrm{Var}(g_{AB})}, \tag{18}$$

where $\overline{V}_{g_{AB}}$ was the mean of 20,000 $V_{g_{AB}}$s obtained under each condition with each $V_{g_{AB}}$computed from Eq. 15, and $\mathrm{Var}(g_{AB})$ was the Monte Carlo variance of 20,000 g_{AB}s computed from Eq. 19,

$$\mathrm{Var}(g_{AB}) = \frac{\sum_1^{20,000}\left(g_{AB} - \overline{g}_{AB}\right)^2}{20,000 - 1}. \tag{19}$$

The Monte Carlo variance, or $\mathrm{Var}(g_{AB})$, was used in Eq. 18 as a proxy for the true variance of g_{AB}. Based on Hoogland and Boomsma (1998), we interpreted $|RBV| < 21\%$ as acceptable and $|RBV| \geq 21\%$ as unacceptable[11]. In addition, $RBV < -21\%$ was interpreted as unacceptable underestimate and $RBV > 21\%$ as unacceptable overestimate.

MSE (mean square error)

MSE measured the precision of g_{AB} as a point estimator. *MSE* is the sum of the squared bias plus variance of g_{AB} which we verified. *MSE* was calculated according to Eq. 20:

$$MSE = \frac{\sum_1^{20,000}\left(g_{AB} - \delta_{AB}\right)^2}{R}, \tag{20}$$

where $\delta_{AB} = 1$ and $R = 20,000$.

To assess the magnitude of *MSE*, we examined *MSE*'s distribution in terms of its mean, median, and maximum. As a point of comparison suggested by Pustejovsky et al. (2014), we compared *MSE*'s mean and median with estimated *MSE*s of Hedges' *g* (Hedges, 1981) obtained from a balanced, two-group experiment with $m \times N$ participants when the population ES is 1. The estimated *MSE* of Hedges'

[10] Pustejovsky et al.'s (2014) supplemental materials defined *RBV* as $\frac{\overline{V}_{g_{AB}}}{\mathrm{Var}(g_{AB})}$. They stated, "We assessed the performance of proposed variance estimators using relative bias; for an effect size estimator *g* [= g_{AB} in the present paper] with associated variance estimator V_g, the relative bias of the variance estimator is the ratio of the expected value of the variance estimator $E(V_g)$ to the true variance of the effect size estimator $\mathrm{Var}(g)$. Relative biases close to one mean that the variance estimator is unbiased." (p. 4).

[11] Hoogland and Boomsma (1998) used 0.10 as the cutoff for an acceptable relative bias in a *SE*. If a sample *SE* is denoted as $\hat{\theta}$ and its parameter as θ, Hoogland and Boomsma's cutoff is expressed as $(\hat{\theta} - \theta)/\theta < 0.1$. Hence, $\hat{\theta}/\theta < 1.1$ or $(\hat{\theta}/\theta)^2 < 1.21$. Therefore, $(\hat{\theta}^2 - \theta^2)/\theta^2 < 0.21$ or 21%.

g with m_g participants in a balanced two-group experiment, when the population ES = 1, is given by Eq. 21:

$$MSE \text{ of Hedges}'\ g = \frac{m_{df}}{(m_{df}-2)\times\left(\frac{m_g}{4}\right)}\times\left[1+\frac{m_g}{4}\right]+\left[1-\frac{2}{c(m_{df})}\right], \quad (21)$$

where $m_{df} = m_g - 2$, $c\left(m_{df}\right) = \frac{\Gamma\left[\frac{m_{df}}{2}\right]}{\sqrt{\frac{m_{df}}{2}}\times\Gamma\left[\frac{m_{df}-1}{2}\right]}$, and Γ = gamma function.

For $m = 3$ and $N = 8$, the *MSE* of Hedges' g is estimated to be 0.212 by setting $m_g = 24$ ($= 3 \times 8$) into Eq. 21. Hence, 0.212 was used as a point of comparison when $m = 3$ and $N = 8$. Similarly, for $m = 4$ and $N = 8$, we specified $m_g = 32$ ($= 4 \times 8$) to obtain a point of comparison = 0.154. For $m = 5$ and $N = 8$, the point of comparison = 0.121. For $m = 6$ and $N = 8$, the point of comparison = 0.099. For $m = 3$, 4, 5, 6 and $N = 16$, the points of comparison = 0.099, 0.073, 0.058, and 0.048, respectively. Additionally, we noted in Table S4 of Pustejovsky et al. (2014), supplemental materials) that maximum *MSE*s were approximately twice as large as the mean *MSE*s across levels of m and N under normal conditions. The overall maximum *MSE* (= 0.664) from Table S4 was approximately four times as large as the overall mean (= 0.167). Extremely large *MSE*s indicated imprecision. And conditions in which these large *MSE*s occurred needed to be identified. Hence, we decided to identify unacceptable *MSE*s as those greater than the 75th percentile of all *MSE*s.

CR (coverage rate of symmetric 95% CI)

Guided by Pustejovsky et al.'s (2014) findings that symmetric CIs of g_{AB} were closer to the nominal level of 95% than noncentral CIs, we constructed the symmetric 95% CI for δ_{AB} using Eq. 22:

$$\text{symmetric 95\% CI for } \delta_{AB} = g_{AB} \pm \sqrt{V_{g_{AB}}} \times t_{0.025,\upsilon}, \quad (22)$$

where $t_{0.025,\upsilon}$ is the critical value from the t distribution with $df = \upsilon$ (Eq. 13), and $V_{g_{AB}}$ is computed from Eq. 15. *CR* was defined as the percentage of the 95% CIs that covered δ_{AB}. We defined an acceptable *CR* to fall between 0.925 (lower bound) and 0.975 (upper bound), according to Algina et al. (2005). A *CR* outside the range of [0.925, 0.975] was deemed unacceptable. In addition, $CR < 0.925$ was interpreted as unacceptable under-coverage and $CR > 0.975$ as unacceptable over-coverage.

Step 6: Analyze the impact of the six factors on four criteria

The impact of the six factors on four criteria (RQ1) was analyzed by four ANOVAs and six plots depicting trends of acceptable and unacceptable criterion values. For each criterion, the ANOVA analyzed the six main effects of Dist, m, N, ρ, λ, and ϕ, plus five two-way interactions of Dist with m, N, ρ, λ, and ϕ, respectively. All effects were treated as fixed. Because each condition yielded one criterion value, the three-way and higher-order interactions were pooled to form the error term in ANOVAs[12]. We defined effects with p-values < 0.05 and eta-squares > 5.9% as having a significant impact on a criterion. An eta-square > 5.9% was labeled by Cohen (1988) as a medium ES.

Step 7: Identify conditions acceptable for MB studies and meta-analysis

To identify conditions in which g_{AB} performed acceptably for MB studies and meta-analysis (RQ2), we applied acceptability standards to the four criteria in each condition. Conditions that yielded all acceptable criteria were identified as acceptable conditions (e.g., Algina et al., 2005; APA, 2020; Hoogland & Boomsma, 1998; Pustejovsky et al., 2014).

Results

Results pertaining to RQ1 are presented first. These include the ANOVA results of the four criteria (*RB*, *RBV*, *MSE*, and *CR*) and trends of acceptable and unacceptable criterion values. The ANOVA results are presented in the section titled "ANOVA results of the four criteria" and trends of criterion values are presented in the section titled "Trends of acceptable and unacceptable criterion values." Results pertaining to RQ2 are presented in the section titled "Acceptable conditions." We summarize all results in "Summary of findings."

ANOVA results of the four criteria

The ANOVA results presented in Table 3 are eta-squares of the six main effects of Dist, m, N, ρ, λ, ϕ and the five two-way interactions of Dist with m, N, ρ, λ, and ϕ on the four criteria. Eta-squares of effects having a significant impact (p-values < 0.05 and eta-squares > 5.9%) are shown in bold. According to Table 3, *RB*'s variance was best explained by all effects with a total eta-square of 91.0%. This was followed by 89.7% of *MSE*'s variance and 83.1% of *CR*'s variance. *RBV*'s variance was least explained with an eta-square of 71.5%.

[12] Alternative three-way ANOVA results that included interactions of Dist with two of the five factors are presented in File 6 at https://osf.io/hsvwu/. The alternative three-way ANOVA identified the same main effects and interactions as the two-way ANOVA, that had significant impacts on the four criteria.

Table 3 Eta-squares (%) of effects on the performance of g_{AB} based on four criteria

Source	*RB*	*RBV*	*MSE*	*CR*
Main effects				
Distribution of data (Dist)	**48.7**	**27.4**	5.6	**23.9**
Number of cases (*m*)	1.6	**22.2**	**34.4**	**9.9**
Number of measurements (*N*)	2.1	**6.0**	3.6	**29.5**
Within-case reliability (ρ)	**12.3**	0.2	**33.9**	1.8
Ratio of variance components (λ)	**12.0**	**6.5**	2.3	0.2
Autocorrelation (Φ)	2.2	3.1	5.9	**9.2**
Interactions				
Dist × *m*	2.7	1.5	1.1	2.1
Dist × *N*	0.1	0.1	0.1	1.2
Dist × ρ	**8.2**	3.3	2.4	1.7
Dist × λ	0.9	1.1	0.3	2.5
Dist × Φ	0.2	0.1	0.1	1.1
Total	91.0	71.5	89.7	83.1

Note. Bolded eta-squares correspond to effects that had a significant impact on a criterion. Bolded eta-squares > 20% are marked by a border. *RB* = relative bias, *RBV* = relative bias of g_{AB}'s variance estimator, *MSE* = mean square error, and *CR* = coverage rate

The Dist factor had a significant impact on *RB*, *RBV*, and *CR*, accounting for most variance of *RB* (explaining 48.7% of its variance) and *RBV* (27.4%). Furthermore, Dist had the second greatest impact on *CR* (23.9%). The *m* factor had a significant impact on *RBV*, *MSE*, and *CR* with the greatest impact on *MSE* (34.4%) and the second greatest impact on *RBV* (22.2%). It is evident that Dist and *m* had greater impact on the four criteria than other factors. The *N* factor had a significant impact on *RBV* and *CR* with the greatest impact on *CR* (29.5%). The ρ factor had a significant impact on *RB* and *MSE*; its impact on *MSE* (33.9%) was the second greatest, only slightly smaller than the greatest impact by *m* (34.4%). The λ factor had a significant impact on *RB* and *RBV*. The ϕ factor had a significant impact on *CR* only.

Regarding two-way interactions of Dist with *m*, *N*, ρ, λ, and ϕ, Dist interacted with ρ in impacting *RB* significantly (8.2% of its variance). Dist did not interact with other factors in impacting *RBV*, *MSE*, or *CR* significantly.

Trends of acceptable and unacceptable criterion values

Based on the ANOVA results, we plot trends of acceptable and unacceptable criterion values as indicators of the performance of g_{AB}. Figures 1 and 2 plot trends of *RB* and *RBV*, respectively. Figures 3 and 4 plot trends of *MSE*. Figure 5 plots trends of *CR* for $N = 8$, whereas Fig. 6 plots trends of *CR* for $N = 16$.

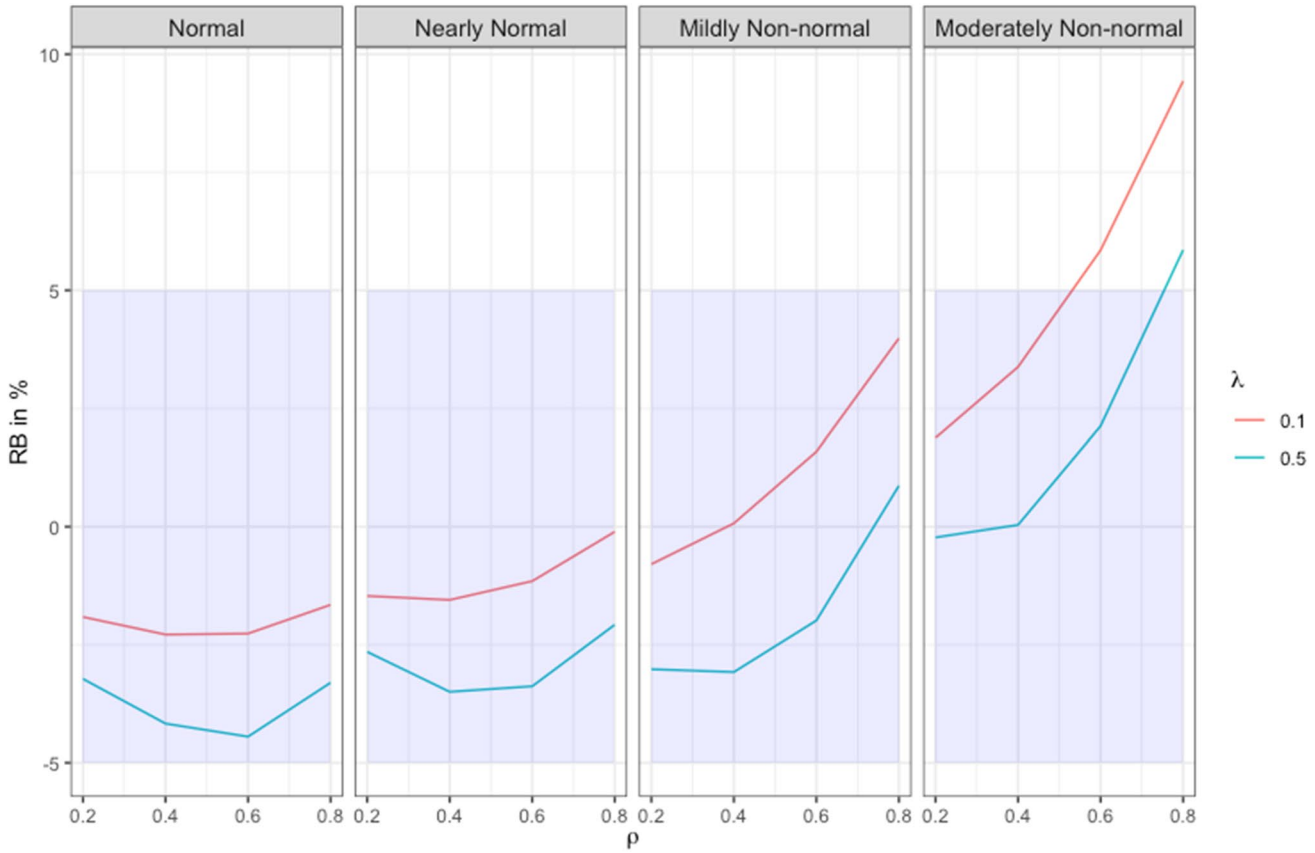

Fig. 1 Effects of data distribution, within-case reliability (ρ), and ratio of variance components (λ) on *RB* (%). *RB*s inside the blue shaded area were acceptable; *RB*s outside the blue shaded area were unacceptable

RB (relative bias)

According to Table 3, *RB* was significantly impacted by Dist, ρ, λ, and the interaction of Dist with ρ. Figure 1 plots *RB* by Dist, ρ, and λ. Acceptable *RB*s are inside blue shaded areas and unacceptable *RB*s are outside blue shaded areas, based on the cutoff for an unacceptable *RB* ($|RB| \geq 5\%$) established in the "Method" section.

According to Fig. 1, *RB*s were all acceptable when Dist was normal, nearly normal, or mildly non-normal. When Dist was moderately non-normal, *RB*s were acceptable when (1) $\rho \leq 0.4$ and $\lambda = 0.1$, or (2) $\rho \leq 0.6$ and $\lambda = 0.5$. Moreover, unacceptable *RB*s were all overestimates.

RB increased as ρ increased and Dist was mildly or moderately non-normal. Given a fixed Dist and ρ, *RB* was smaller when $\lambda = 0.5$ than when $\lambda = 0.1$.

RBV (relative bias of g_{AB}'s variance estimator)

According to Table 3, *RBV* was significantly impacted by Dist, *m*, *N*, and λ. Figure 2 plots *RBV* by Dist, *m*, *N*, and λ. Acceptable *RBV*s are inside blue shaded areas and unacceptable *RBV*s are outside blue shaded areas, based on the cutoff for an unacceptable *RBV* ($|RBV| \geq 21\%$) established in the "Method" section.

According to Fig. 2, when Dist was normal, *RBV*s were acceptable only when $m = 6$ and $N = 16$. When Dist was nearly normal, *RBV*s were acceptable when $m = 6$ and $N = 16$, or when $m = 5$, $N = 16$, and $\lambda = 0.1$. When Dist was mildly non-normal, *RBV*s were acceptable if (1) $\lambda = 0.1$, (2) $m \geq 5$, $N = 8$, $\lambda = 0.5$, or (3) $m \geq 4$, $N = 16$, and $\lambda = 0.5$. When Dist was moderately non-normal, *RBV*s were acceptable if (1) $m \geq 4$, $N = 8$, and $\lambda = 0.1$, (2) $m \geq 5$, $N = 8$, and $\lambda = 0.5$, (3) $N = 16$ and $\lambda = 0.1$, or (4) $m \geq 4$, $N = 16$, and $\lambda = 0.5$. Moreover, unacceptable *RBV*s were all overestimates.

*RBV*s obtained from normal and nearly normal Dists were similar; they were noticeably larger than those obtained from mildly or moderately non-normal Dist. *RBV* decreased as *m* increased from 3 to 6, *N* increased from 8 to 16, or λ decreased from 0.5 to 0.1.

MSE (mean square error)

Table 4 presents means, medians, and maximums of g_{AB}'s *MSE* and points of comparison based on estimated *MSE*s of Hedges' *g*. It was evident that means and medians of g_{AB}'s

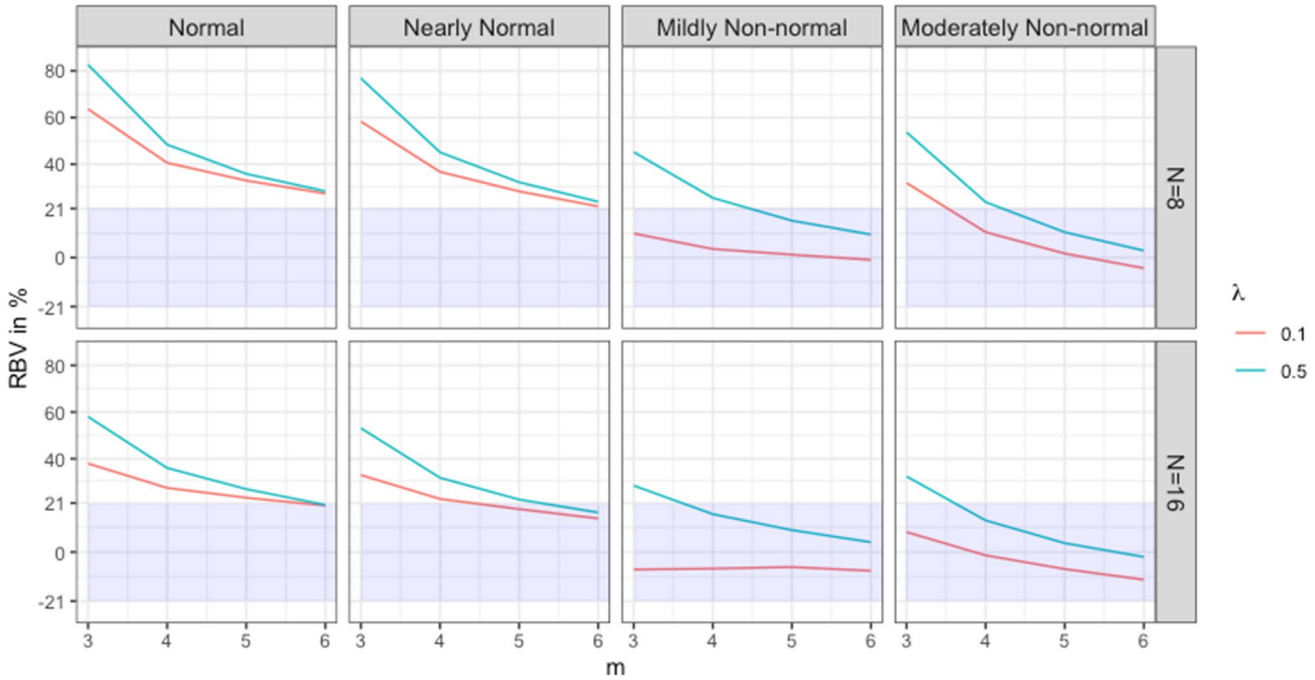

Fig. 2 Effects of data distribution, number of cases (*m*), number of measurements (*N*), and ratio of variance components (λ) on *RBV* (%). *RBV*s inside the blue shaded areas were acceptable; *RBV*s outside the blue shaded areas were unacceptable.

MSE were noticeably larger than *MSE*s of Hedges' *g*. When $N=8$, g_{AB}'s mean *MSE*s were approximately 1.5 times larger than *MSE*s of Hedges' *g*. When $N=16$, g_{AB}'s mean *MSE*s were approximately 2.5 times larger than *MSE*s of Hedges' *g*. Similar to Pustejovsky et al.'s (2014) findings, extremely large *MSE*s were uncovered in this study. Maximum *MSE*s were approximately twice as large as mean *MSE*s across levels of *m* and *N*. The overall maximum *MSE* (=0.808) was approximately four times as large as the overall mean (=0.194). Thus, we decided that it was justified to use the 75th percentile (=0.241) of all *MSE*s as a cutoff to identify unacceptable *MSE*s.

According to Table 3, *MSE* was significantly impacted by *m* and ρ. Figure 3 plots *MSE* by *m*, ρ, and Dist. Acceptable *MSE*s are inside the blue shaded areas and unacceptable *MSE*s are outside the blue shaded areas, based on the cutoff for an unacceptable *MSE* (*MSE* > 0.241).

According to Fig. 3, *MSE* decreased as *m* increased for a fixed ρ, regardless of data distribution. At a fixed *m*, *MSE* increased as ρ increased, again regardless of data distribution. When $m=6$, all *MSE*s were acceptable. When $m=4$ or 5, all *MSE*s were acceptable, except when $\rho=0.8$ and Dist was mildly or moderately non-normal. When $m=3$, *MSE*s were acceptable only when $\rho=0.2$, or when $\rho=0.4$ and Dist was normal or nearly normal.

Figure 4 presents boxplots of *MSE*s for each combination of *m* and ρ across levels of Dist, *N*, λ, and ϕ. Three reference lines are overlaid corresponding to the 95th percentile (=0.401), 75th percentile (=0.241), and median (=0.170) of the 1792 *MSE*s, respectively. According to Fig. 4, when $m=3$ (the smallest) and $\rho=0.8$ (the largest), all *MSE*s were larger than 0.241 ($=P_{75}$). When $m=3$ and $\rho=0.6$, or when $m=4$ and $\rho=0.8$, more than 75% of *MSE*s were larger than 0.241 ($=P_{75}$). Under all other combinations of *m* and ρ, more than 50% *MSE*s were smaller than 0.241 ($=P_{75}$). The exact P_{25}, median, mean, P_{75}, and P_{95} of each boxplot shown in Fig. 4 are presented in File 7.

CR (coverage rate of symmetric 95% CI)

According to Table 3, *CR* was significantly impacted by Dist, *m*, *N*, and ϕ. Figure 5 plots *CR* by Dist, *m,* and ϕ for $N=8$, whereas Fig. 6 plots *CR* by the same factors for $N=16$. Acceptable *CR*s are inside the blue shaded areas and unacceptable *CR*s are outside the blue shaded areas, based on the cutoff for unacceptable *CR* (*CR* < 0.925 or > 0.975) established in the "Method" section.

When $N=8$ and $m=5$ or 6, *CR*s were mostly acceptable across the four Dists. As *m* decreased from 5, 4 to 3, *CR*s became increasingly unacceptable and overcovering, especially under normal and nearly normal distributions. When $m=3$ or 4 under mildly and moderately non-normal Dists, *CR*s were mostly unacceptable when $\phi \geq 0$.

When $N=16$, *CR*s were mostly acceptable across the four Dists. A few unacceptable overcovering *CR*s were found when (1) $m=3$ or 4 and Dist was normal or nearly normal,

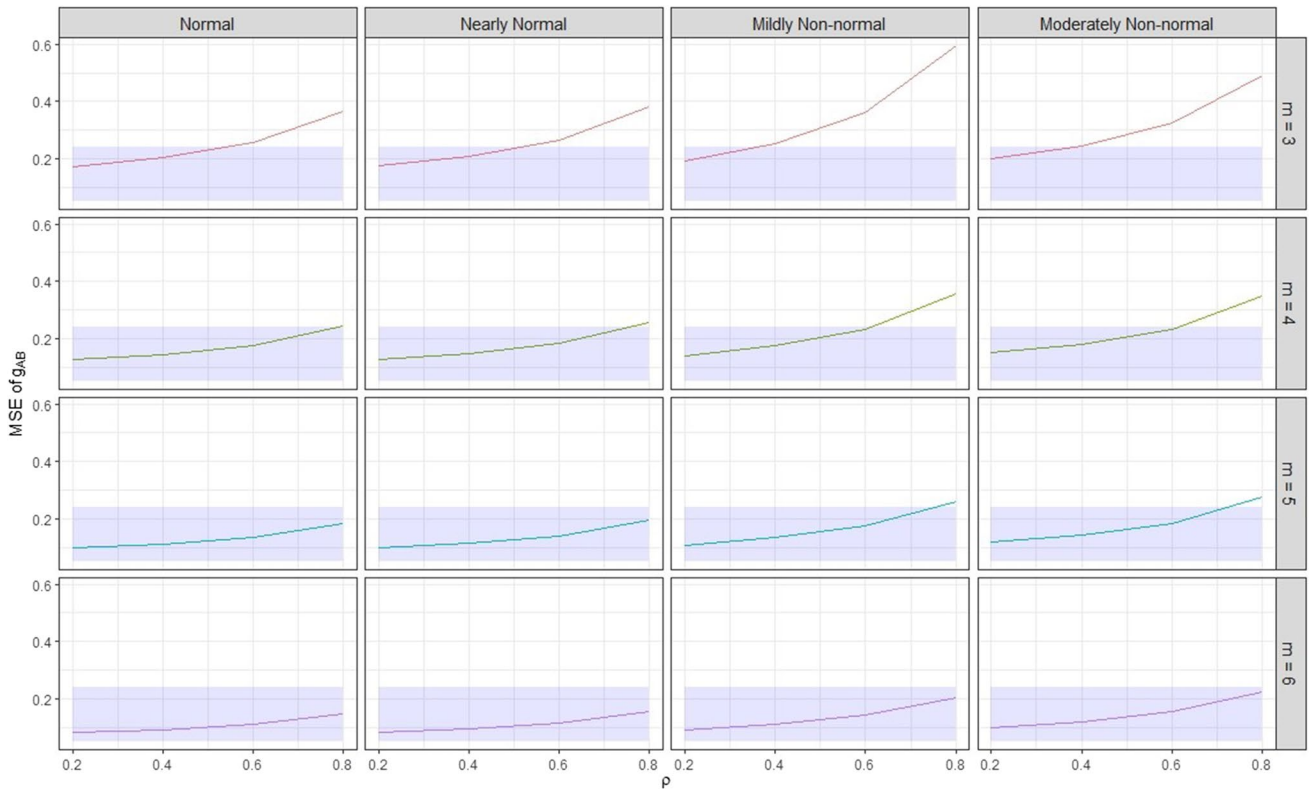

Fig. 3 Effects of number of cases (*m*) and within-case reliability (ρ) on *MSE* by data distribution. *MSE*s inside the blue shaded area were acceptable; *MSE*s outside the blue shaded area were unacceptable

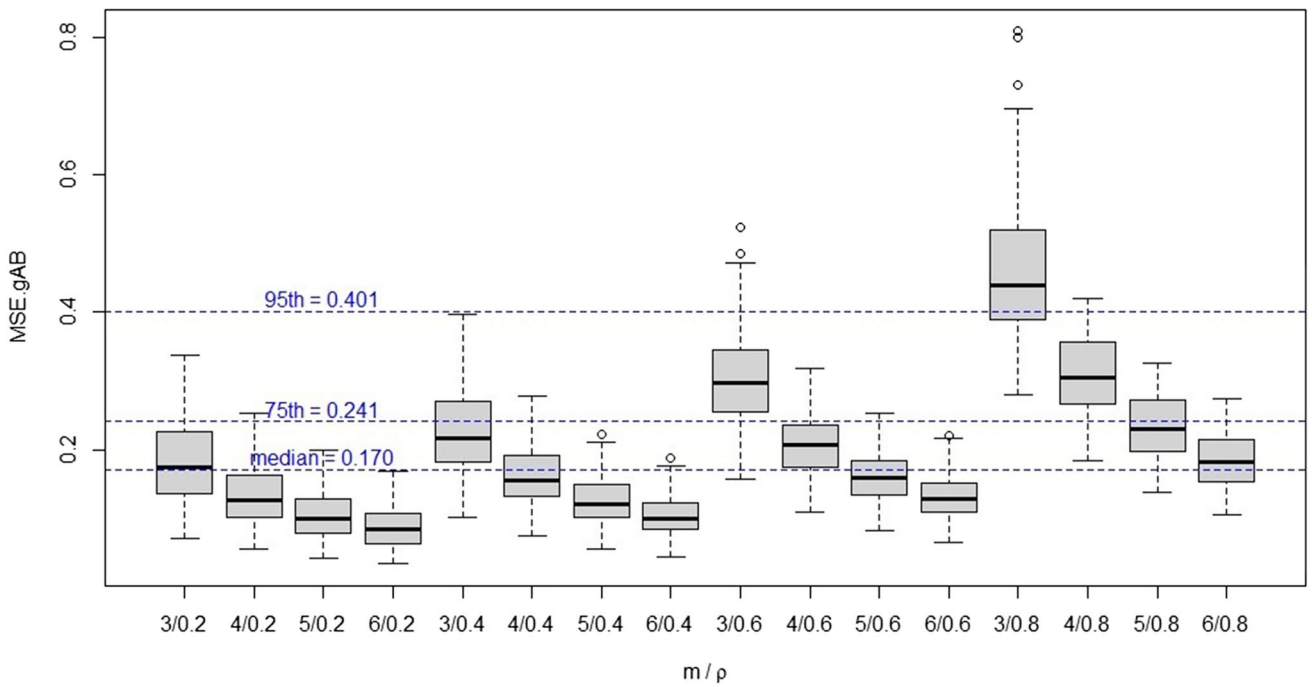

Fig. 4 Boxplots of *MSE*s for combinations of number of cases (*m*) and within-case reliability (ρ)

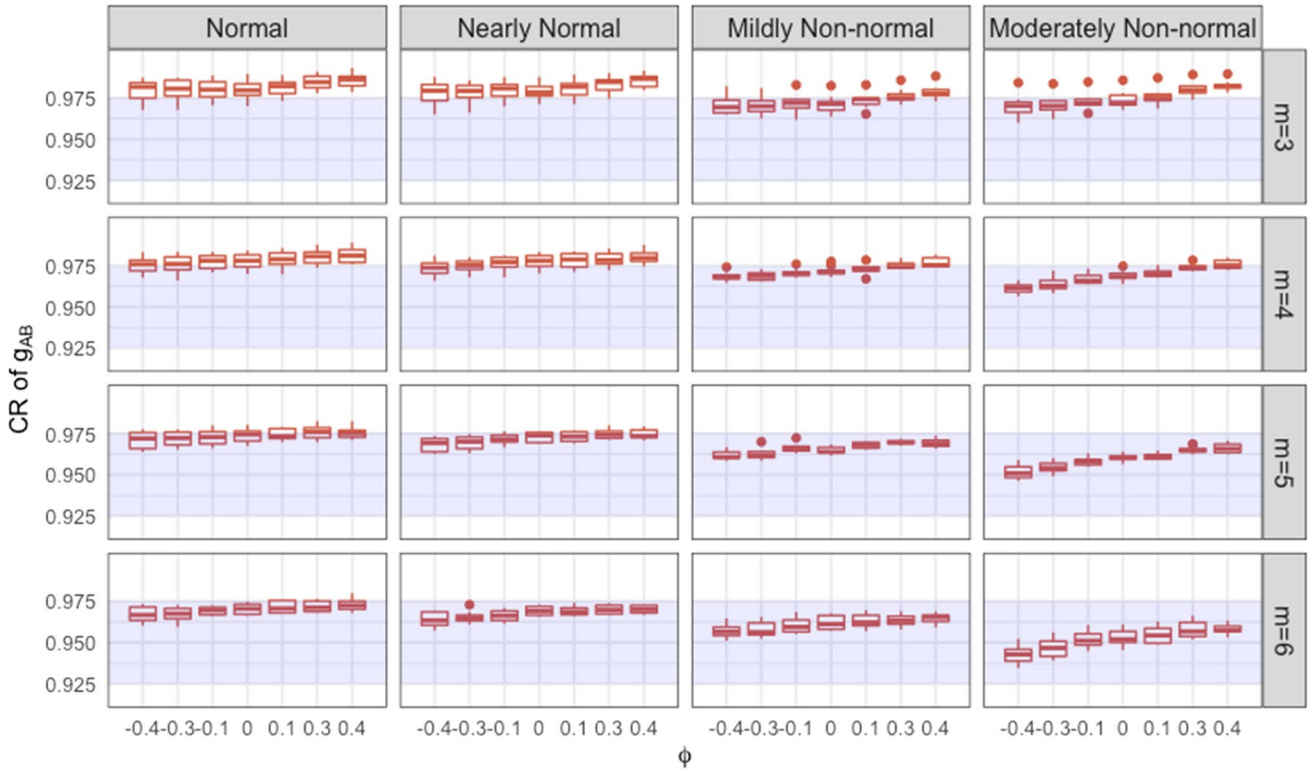

Fig. 5 Effects of data distribution, number of cases (m), and autocorrelation (ϕ) on *CR* when $N = 8$. *CR*s inside blue shaded areas were acceptable; *CR*s outside the blue shaded areas were unacceptable

or (2) $m = 3$, $\phi \geq 0$, and Dist was mildly or moderately non-normal. A few unacceptable undercovering *CR*s were identified when $m = 5$ or 6, $\phi \leq -0.3$, and Dist was moderately non-normal.

Acceptable conditions

To answer RQ2, we integrated findings across four criteria in Tables 5, 6, 7 and 8. Specifically, Tables 5 and 6 present acceptable conditions for $\lambda = 0.1$ and $N = 8$ or 16, respectively. Tables 7 and 8 present acceptable conditions for $\lambda = 0.5$ and $N = 8$ or 16, respectively. A circle (O) in Tables 5, 6, 7 and 8 indicates that g_{AB} performed acceptably on all four criteria in that condition; hence, g_{AB} was acceptable for MB studies and meta-analysis in that condition (e.g., Algina et al., 2005; APA, 2020; Hoogland & Boomsma, 1998; Pustejovsky et al., 2014). Actual values of the four criteria are shown in File 5 at https://osf.io/hsvwu/.

According to Tables 5, 6, 7 and 8, the number of acceptable conditions marked by O ranged from 55 (normal), 90 (nearly normal), 223 (moderately non-normal), to 233 (mildly non-normal). g_{AB}'s performance was acceptable in far more mildly or moderately non-normal conditions than normal or nearly normal conditions. At fixed λ and N, normal and nearly normal Dists yielded similar patterns, and mildly and moderately non-normal Dists yielded similar patterns.

The number of acceptable conditions increased as m or N increased. When m increased from 3, 4, 5 to 6, the number of acceptable conditions increased from 51, 113, 182, to 255, respectively. When N increased from 8 to 16, the number of acceptable conditions increased from 220 to 381.

In general, the number of acceptable conditions decreased as λ, ρ or ϕ increased. When λ increased from 0.1 to 0.5, the number of acceptable conditions decreased from 320 to 281. When ρ increased from 0.2, 0.4, 0.6 to 0.8, the number of acceptable conditions in general decreased from 172, 180, 157, to 92, respectively. When ϕ increased from −0.4, −0.3, −0.1, 0, 0.1, 0.3 to 0.4, the number of acceptable conditions gradually decreased from 110, 106, 87, 82, 81, 68, to 67, respectively.

According to Table 5, when $\lambda = 0.1$ and $N = 8$, most of the acceptable conditions were associated with (1) $\rho = 0.2$ or 0.4, $m \geq 4$, and mildly or moderately non-normal Dist, (2) $\rho = 0.6$, $m \geq 5$, and mildly non-normal Dist, or (3) $\rho = 0.8$, $m = 6$, and normal, nearly normal, or mildly non-normal Dist. According to Table 6, when $\lambda = 0.1$ and $N = 16$, most of the acceptable conditions were associated with (1) $\rho = 0.2$ or 0.4, and mildly or moderately non-normal Dist, (2) $\rho = 0.6$,

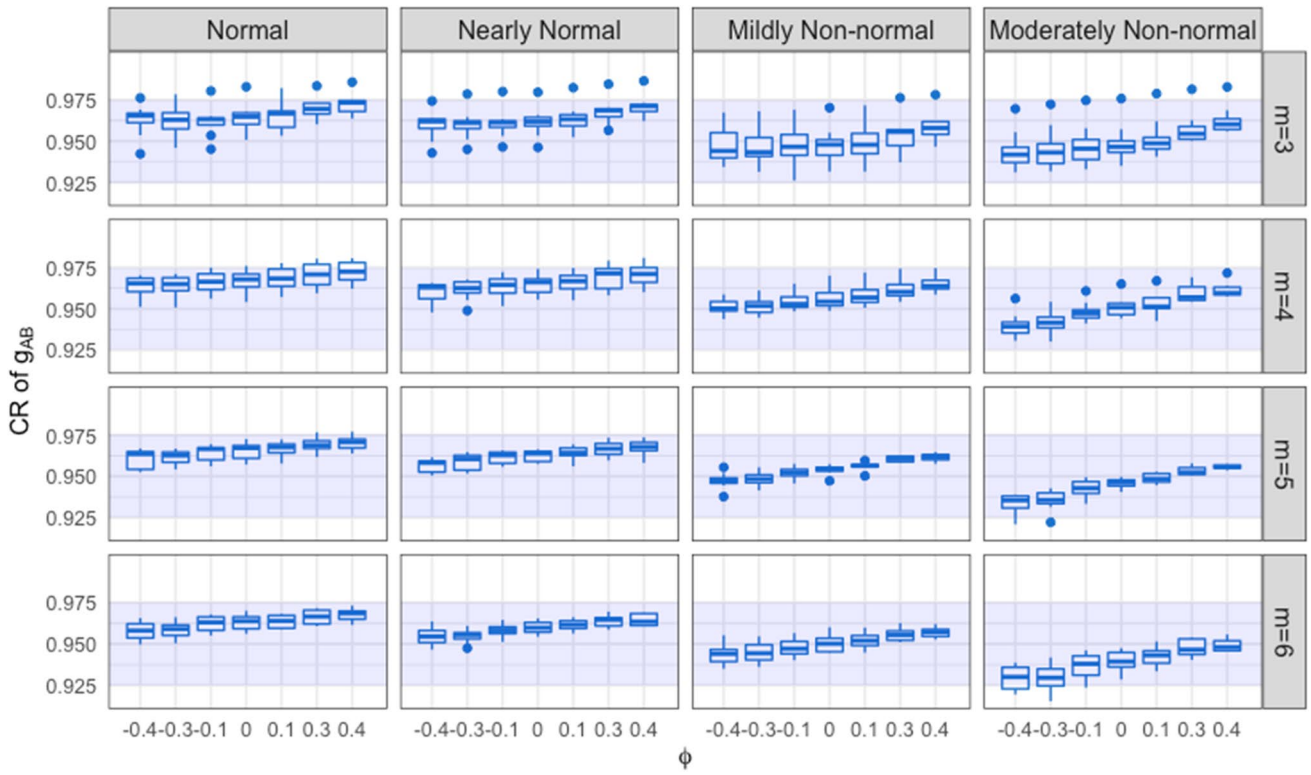

Fig. 6 Effects of data distribution, number of cases (m), and autocorrelation (ϕ) on *CR* when $N = 16$. *CR*s inside blue shaded areas were acceptable; *CR*s outside the blue shaded areas were unacceptable

$m \geq 4$, and normal, nearly normal, or mildly non-normal Dist, or (3) $\rho = 0.8$, $m \geq 5$, and normal or nearly normal Dist.

According to Table 7, when $\lambda = 0.5$ and $N = 8$, most of the acceptable conditions were associated with (1) $\rho = 0.2$, $m \geq 4$, and moderately non-normal Dist, or (2) $\rho = 0.4$ or 0.6, $m \geq 5$, and mildly or moderately non-normal Dist. According to Table 8, when $\lambda = 0.5$ and $N = 16$, most of the acceptable conditions were associated with (1) $\rho = 0.2$, and moderately non-normal Dist, (2) $\rho = 0.2$, $m = 3$ or 6, and mildly non-normal Dist, (3) $\rho = 0.4$ or 0.6 , $m \geq 4$, and mildly or moderately non-normal Dist, (4) $\rho = 0.4$, 0.6, or 0.8, $m = 6$, and normal, nearly normal, or mildly non-normal Dist, or (5) $\rho = 0.8$, $m = 5$, and nearly normal Dist.

Summary of findings

Results presented above indicated that g_{AB} as a point estimator was fairly unbiased even under non-normal distributions. g_{AB}'s variance was generally overestimated and its 95% CI was over-covered, especially when data distribution was normal or nearly normal combined with $m = 3$ or 4, and $N = 8$. The imprecision of g_{AB}, as measured by *MSE*, was quite large when $m = 3$ or 4 and $\rho = 0.6$ or 0.8 across the four distributions.

Indeed, data distribution played a vital role in impacting g_{AB}'s performance for MB studies and meta-analysis.

Table 4 Mean, median, and maximum *MSE*s of g_{AB} and estimated *MSE*s of Hedges' g as points of comparison derived from Eq. 21

m	$N = 8$				$N = 16$			
	Mean	Median	Maximum	Points of comparison	Mean	Median	Maximum	Points of comparison
3	0.326	0.308	0.808	0.212	0.259	0.232	0.667	0.099
4	0.222	0.209	0.419	0.154	0.182	0.165	0.383	0.073
5	0.171	0.162	0.325	0.121	0.140	0.128	0.313	0.058
6	0.140	0.133	0.275	0.099	0.113	0.104	0.256	0.048

Table 5 Acceptable conditions when $\lambda = 0.1$ and $N = 8$ for MB studies and meta-analysis

		Normal				Nearly normal				Mildly non-normal				Moderately non-normal			
ρ	ϕ	3	4	5	6	3	4	5	6	3	4	5	6	3	4	5	6
0.2	−0.4											O	O	O	O	O	O
	−0.3												O	O	O	O	O
	−0.1										O		O		O	O	O
	0											O	O		O	O	O
	0.1										O	O	O		O	O	O
	0.3										O	O	O		O	O	O
	0.4											O	O		O	O	O
0.4	−0.4									O	O	O	O	O	O	O	O
	−0.3									O	O	O	O	O	O	O	O
	−0.1										O	O	O		O	O	O
	0										O	O	O		O	O	O
	0.1										O	O	O		O	O	O
	0.3											O	O			O	O
	0.4											O	O			O	O
0.6	−0.4				O			O	O		O	O	O		O		O
	−0.3				O				O		O	O	O			O	
	−0.1										O	O	O				
	0											O	O			O	
	0.1											O	O				O
	0.3											O	O				
	0.4											O	O			O	O
0.8	−0.4				O			O	O								
	−0.3				O			O	O								
	−0.1				O				O			O	O				
	0				O				O				O				
	0.1								O				O				
	0.3								O				O				
	0.4								O				O				

Note. 3, 4, 5, 6 in the header row are the number of cases

g_{AB} performed far better under mildly and moderately non-normal distributions than under normal and nearly normal distributions. This was because more *RBV*s were acceptable under mildly or moderately non-normal distribution than under normal or nearly normal distribution (see Fig. 2). Additionally, data distribution interacted with ρ in impacting the performance of g_{AB} significantly. Under normal or nearly normal distribution, g_{AB} performed more acceptably when $\rho = 0.6$ or 0.8 than when $\rho = 0.2$ or 0.4. Under mildly or moderately non-normal distribution, g_{AB} performed more acceptably when $\rho = 0.2$ or 0.4 than when $\rho = 0.8$. When $\rho = 0.6$ and $\lambda = 0.5$, g_{AB} performed equally acceptably under mildly and moderately distributions. When $\rho = 0.6$ and $\lambda = 0.1$, g_{AB} performed more acceptably under the mildly non-normal than under the moderately non-normal distribution. The negative impact of ρ on g_{AB} under any data distribution was mitigated by doubling N from 8 to 16 and/or by increasing m from 3 to 6.

Discussion

The *What Works Clearinghouse Procedures and Standards Handbook, Version 5.0* (WWC, 2022) recommends g_{AB} as a D-CES (1) to gauge an intervention effect in SCED studies, and also (2) to synthesize findings from SCED and group studies in meta-analysis. As an estimator of its parameter δ_{AB}, g_{AB}'s interpretation is conditioned on its assumptions. Yet there have been no published studies that examined g_{AB}'s normality assumption. The present study aimed to investigate the impact of data distributions on the performance of g_{AB} by expanding Study 2 of Pustejovsky et al. (2014) to non-normal data. Our study

Table 6 Acceptable conditions when $\lambda = 0.1$ and $N = 16$ for MB studies and meta-analysis

		Normal				Nearly normal				Mildly non-normal				Moderately non-normal			
ρ	ϕ	3	4	5	6	3	4	5	6	3	4	5	6	3	4	5	6
0.2	−0.4									O	O	O	O	O	O	O	O
	−0.3									O	O	O	O	O	O	O	O
	−0.1									O	O	O	O	O	O	O	O
	0									O	O	O	O	O	O	O	O
	0.1									O	O	O	O	O	O	O	O
	0.3									O	O	O	O	O	O	O	O
	0.4									O	O	O	O	O	O	O	O
0.4	−0.4				O		O	O	O	O	O	O	O	O	O	O	
	−0.3							O	O	O	O	O	O	O	O	O	O
	−0.1								O	O	O	O	O		O	O	O
	0									O	O	O	O		O	O	O
	0.1									O	O	O	O		O	O	O
	0.3										O	O	O		O	O	O
	0.4										O	O	O		O	O	O
0.6	−0.4		O	O	O	O	O	O	O								
	−0.3		O	O	O		O	O	O		O	O					
	−0.1				O		O	O	O		O	O	O				
	0			O	O		O	O	O		O	O	O				
	0.1				O			O	O		O	O	O				
	0.3								O		O	O	O				O
	0.4								O		O	O	O				O
0.8	−0.4		O	O	O		O	O	O								
	−0.3		O	O	O		O	O	O								
	−0.1			O	O		O	O	O								
	0			O	O		O	O	O								
	0.1			O	O		O	O	O								
	0.3				O			O	O			O	O				
	0.4				O			O	O								

Note. 3, 4, 5, 6 in the header row are the number of cases

applied the same REML method to compute g_{AB} and the same four criteria to evaluate the performance of g_{AB}, as Pustejovsky et al. (2014) did. The four criteria were: relative bias (*RB*), relative bias of g_{AB}'s variance estimator (*RBV*), mean square error (*MSE*), and coverage rate of symmetric 95% CI (*CR*). The present study differed from Pustejovsky et al. (2014) in two aspects. First, we analyzed converged results, whereas Pustejovsky et al. (2014) analyzed both converged and non-converged results. Second, we specified cutoffs for acceptable and unacceptable *RB*, *RBV*, *MSE*, and *CR*, whereas Pustejovsky et al. (2014) did not establish such standards.

Two research questions (RQ1 and RQ2) were raised concerning the impact of data distributions on the performance of g_{AB}. RQ1 explored the extent to which data distribution (Dist), the number of cases (m), the number of measurements (N), within-case reliability (ρ), ratio of variance components (λ), autocorrelation (ϕ) and the interactions of Dist with m, N, ρ, λ, and ϕ impacted each of the four criteria. Dist was manipulated in this study to range from normal, nearly normal, mildly non-normal to moderately non-normal. The other five factors (m, N, ρ, λ, and ϕ) were manipulated identically or similarly as in Pustejovsky et al. (2014). We answered RQ1 by analyzing the impact of the six main effects and five interactions on each criterion in an ANOVA framework (Table 3) and by plotting trends of acceptable and unacceptable *RB* (Fig. 1), *RBV* (Fig. 2), *MSE* (Figs. 3 and 4), and *CR* (Figs. 5 and 6).

We answered RQ2 by applying acceptability standards to all four criteria in each condition (Tables 5, 6, 7 and 8). The suitability of g_{AB} for MB studies and meta-analysis (RQ2) was answered by identifying conditions that yielded acceptable results on all four criteria. Our results indicated that g_{AB} as a point estimator was fairly unbiased even under non-normal distributions. Yet g_{AB}'s variance was generally

Table 7 Acceptable conditions when $\lambda = 0.5$ and $N = 8$ for MB studies and meta-analysis

		Normal				Nearly normal				Mildly non-normal				Moderately non-normal			
ρ	ϕ	3	4	5	6	3	4	5	6	3	4	5	6	3	4	5	6
0.2	−0.4														O	O	O
	−0.3													O	O	O	O
	−0.1														O	O	O
	0														O	O	O
	0.1												O		O	O	O
	0.3											O	O		O	O	O
	0.4											O	O			O	O
0.4	−0.4												O		O	O	O
	−0.3												O		O	O	O
	−0.1												O			O	O
	0												O			O	O
	0.1											O	O			O	O
	0.3											O	O			O	O
	0.4											O	O			O	O
0.6	−0.4				O				O			O	O		O	O	O
	−0.3				O				O			O	O			O	O
	−0.1											O	O			O	O
	0											O	O			O	O
	0.1											O	O			O	O
	0.3											O	O				O
	0.4												O				O
0.8	−0.4				O			O	O				O				
	−0.3				O				O				O				
	−0.1								O				O				
	0												O				
	0.1																
	0.3																
	0.4																

Note. 3, 4, 5, 6 in the header row are the number of cases

overestimated, and its symmetric 95% CI was over-covered, especially when data distribution was normal or nearly normal combined with $m = 3$ or 4, and $N = 8$. The imprecision of g_{AB}, as measured by *MSE*, was a concern when $m = 3$ or 4 and $\rho = 0.6$ or 0.8, regardless of data distribution.

Under normal or nearly normal data distribution, bias in variance estimates of g_{AB} was mostly unacceptable. In contrast, more *RBV*s were acceptable under mildly or moderately non-normal distribution than under normal or nearly normal distribution. Consequently, g_{AB} was suitable for MB studies and meta-analysis in more conditions under mildly or moderately non-normal distribution than under normal or nearly normal distribution. It may seem counterintuitive as to why more *RBV*s were unacceptable under normal and nearly normal Dists than under mildly and moderately non-normal Dists. One explanation is given by Eq. 15 on the estimated variance of g_{AB}, or $V_{g_{AB}}$. Though Eq. 15 is derived under normality (Pustejovsky et al., 2014), its approximation to the true variance of g_{AB} is poor when m and N are small due to asymptotic normality of the REML method (see Footnote 2). Our results revealed that the largest m (= 6) and N (= 16) in this study were not large enough to yield acceptable estimates of the variance of g_{AB} under normal and nearly normal Dists[13]. Our results also showed that Eq. 15 yielded more biased estimates of g_{AB}'s variance under small m (= 3) and N (= 8) than under large m (= 6) and N (= 16). These findings are consistent with the mathematical derivation of Eq. 15 under the normality assumption. Given the substantial relative bias in g_{AB}'s variance estimator in the present study and in Pustejovsky et al. (2014), Eq. 15 needs to be improved in future research. We offer a few viable

[13] The fact that more *RBV*s were acceptable under mildly and moderately non-normal conditions may be explained by the compensation of non-normality for REML's tendency to overestimate the true variance of g_{AB} under normal and nearly normal conditions.

Table 8 Acceptable conditions when $\lambda = 0.5$ and $N = 16$ for MB studies and meta-analysis

		Normal				Nearly normal				Mildly non-normal				Moderately non-normal			
ρ	ϕ	3	4	5	6	3	4	5	6	3	4	5	6	3	4	5	6
0.2	−0.4								O	O	O	O	O	O	O	O	O
	−0.3								O	O	O	O	O	O	O	O	O
	−0.1									O			O	O	O	O	O
	0					O				O	O		O	O	O	O	O
	0.1									O			O	O	O	O	O
	0.3									O			O	O	O	O	O
	0.4									O	O	O	O	O	O	O	O
0.4	−0.4			O	O			O	O	O	O	O	O	O	O	O	O
	−0.3				O			O	O	O	O	O	O	O	O	O	O
	−0.1				O				O	O	O	O	O		O	O	O
	0								O			O	O		O	O	O
	0.1								O	O		O	O	O	O	O	O
	0.3											O	O		O	O	O
	0.4										O	O	O		O	O	O
0.6	−0.4			O	O			O	O		O	O	O		O	O	
	−0.3			O	O			O	O		O	O	O		O	O	O
	−0.1				O			O	O		O	O	O		O	O	O
	0				O				O		O	O	O		O	O	O
	0.1				O				O		O	O	O		O	O	O
	0.3											O	O			O	O
	0.4											O	O			O	O
0.8	−0.4			O	O			O	O				O				
	−0.3			O	O			O	O				O				
	−0.1			O	O			O	O				O				
	0				O			O	O				O				
	0.1				O			O	O				O				
	0.3				O				O				O				
	0.4								O				O				

Note. 3, 4, 5, 6 in the header row are the number of cases

alternative methods under "Limitations and future research directions."

Our results revealed a complex and joint impact of data distribution, along with number of cases, number of measurements, ratio of variance components, within-case reliability, and autocorrelation on the suitability of g_{AB} for MB studies and meta-analysis. According to the ANOVA results, data distribution contributed to approximately 49% of variance in *RB* and 25% of variance in both *RBV* and *CR*. Furthermore, data distribution interacted with within-case reliability in impacting 8% of variance in *RB*. Number of cases and within-case reliability each contributed to 34% of variance in *MSE*. Among the four criteria, *RB* was most accounted for (91%) and *RBV* was least accounted for (72%) by all effects combined.

Our findings showed that the performance of g_{AB}, as assessed by *RB*, *MSE*, *RBV*, and *CR*, depended on the distribution of data and five data features investigated in the present study. Of the six data features, the number of cases and measurements are within the control of a SCED researcher or interventionist, whereas data distribution, within-case reliability, ratio of variance components, and autocorrelation are not. In fact, our results demonstrated a mitigating effect of increased cases and measurements on the negative impact due to increased ratio of variance components, within-case reliability, and first-order autocorrelation on the performance of g_{AB}. The mitigating effect of increased m or N on g_{AB} was particularly evident for normal and nearly normal distributions.

Furthermore, large number of cases and measurements are required by the REML method to yield acceptable estimates of g_{AB}'s variance. Therefore, SCED researchers and interventionists should be encouraged to design a study with sufficiently large number of cases and measurements. In light of our findings, we offer general recommendations of g_{AB} for primary MB studies and meta-analysis.

The preamble to our recommendations is that each SCED study, for which g_{AB} is applicable, had been well constructed, conducted, and documented (Chen et al., 2020; Kratochwill et al., 2013, 2021; Peng et al., 2013).

Our recommendations

We recommend g_{AB} for primary MB studies and meta-analysis when each study with at least 16 measurements, meets one of the following two conditions:

(1) $m = 6$, the within-case reliability is 0.6 or 0.8, and the shape, skewness, and kurtosis of the data distribution are similar to the normal or nearly normal distribution investigated in this study;
(2) $m \geq 4$, the within-case reliability is 0.2, 0.4, or 0.6, and the shape, skewness, and kurtosis of the data distribution are similar to the mildly or moderately non-normal distribution investigated in this study.

The shape, skewness and kurtosis may be determined from raw data and compared to the four data distributions manipulated in the current study (see File 1 at https://osf.io/hsvwu/ for the four marginal distributions). Alternatively, the shape, skewness and kurtosis of empirical data distributions concerning a specific intervention or outcome measure may be conjectured from review studies, such as Joo (2017), Shadish et al. (2014), and Solomon (2014). Likewise, the within-case reliability of a study's data may be computed from the ratio of the between-case variance in levels over the sum of the within-case variance plus the between-case variance in levels, and then compared to the four levels manipulated in the current study.

In empirical SCED studies, there may well be exceptions to the conditions recommended above when g_{AB} performs satisfactorily. When applied researchers find their data to not meet the recommended conditions, a cautionary note should be added for interpreting the magnitude of g_{AB}. Practitioners are advised to consider context effects of an intervention holistically when interpreting g_{AB} and its reasonableness.

Limitations and future research directions

As with any simulation study, findings and recommendations presented here are based on specific manipulations of the six factors investigated and acceptable standards established in the present study. Of the six factors, the number of cases (m), the within-case reliability (ρ), and the autocorrelation (ϕ) were well represented by levels commonly found in empirical MB studies. The number of measurements ($N = 8$ or 16) and the ratio of variance components ($\lambda = 0.1$ or 0.5) were represented by only two levels.

The manipulation of data distributions in the present simulation study ranged from normal to moderately non-normal. All four distributions were unimodal, either mesokurtic or leptokurtic. And the two non-normal distributions were both positively skewed. Future studies can examine performance of g_{AB} using discrete or count/frequency data with one or more modes, or different distributions at Level-1 and Level-2.

We fitted MB2 to normal and non-normal data. Hence, the impact of non-normality on g_{AB} has not been investigated under MB1, MB3, MB4, or MB5 (Pustejovsky et al., 2014). Simulation studies conducted under models other than MB2 should yield useful information to inform SCED researchers and interventionists about g_{AB} in broader contexts. Studies on the impact of model misspecification should facilitate our understanding of the application of g_{AB}.

Relative bias in g_{AB}'s variance estimator was substantial in the present study and in Pustejovsky et al. (2014). Such relative bias has impeded g_{AB}'s utility in meta-analysis. In the context of hierarchical modeling of SCED data, alternative methods, such as the Bayesian method (Baek et al., 2020; Joo & Ferron, 2019), have been proposed to improve the variance estimation of a random effect. However, biased variance estimation remains an unresolved issue in meta-analysis of SCED studies.

In sum, future studies should consider: (a) generating data from non-normal distributions not considered in the present study (e.g., platykurtic, discrete, count/frequency, bimodal/multimodal); (b) simulating data from different distributions for Level-1 and Level-2 errors separately; (c) increasing the number of cases > 6 and the number of measurements ≥ 20 to improve the convergence of REML estimates (see File 8 for non-convergence rates); (d) fitting a model different from MB2 to data; and (e) developing Bayesian or robust estimators of g_{AB}'s variance (e.g., Chen & Pustejovsky, 2022; Park & Beretvas, 2019; Tipton, 2015; Verbeke & Lesaffre, 1996; Yuan & Bentler, 2002), or applying the Box-Cox or square root transformation of g_{AB}'s variance (Man et al., 2022).

Concluding remarks

The g_{AB} statistic has been endorsed as a D-CES by the *What Works Clearinghouse Procedures and Standards Handbook, Version 5.0* (WWC, 2022) to gauge an intervention effect in a SCED study and to synthesis findings from multiple SCED studies, or across SCED and group studies. Such an endorsement will surely increase the reporting of g_{AB} in published reports. Thus, it is crucial to examine the performance of g_{AB} under non-normal distributions with different data features in order to render appropriate reporting and interpretation of g_{AB}. Findings from our study highlight the importance of data distributions and features in determining the suitability of g_{AB} for primary MB studies and meta-analysis.

Based on g_{AB}'s definition and its REML estimation method, it is worth pointing out several issues associated with g_{AB}'s applicability in SCED contexts. First, g_{AB} is applicable to MB designs across three or more cases of the same behavior, as previously mentioned. g_{AB} is equally applicable to AB^k designs with at least three cases of the same behavior. Second, as a study-level ES index, g_{AB} does not permit the examination of factors that may vary between cases within a study, as in moderator analyses (Kratochwill et al., 2021). Third, g_{AB}'s computation and interpretation depend on a model. Such a model may be misspecified, or overly simplistic (Maggin et al., 2022; Valentine et al., 2016)[14], and the model assumptions (e.g., normality, equally spaced measurements, no trend) may be untenable (Maggin et al., 2022) or non-robust. Fourth, its applicability beyond the "starting point," or MB1, recommended by the *What Works Clearinghouse Procedures and Standards Handbook, Version 5.0* (WWC, 2022) requires extensive research, especially under non-normal distributions with small number of cases and measurements.

In addition to issues discussed above, Kratochwill et al. (2021) raised concerns about the compatibility of constructs underlying outcomes measured in SCED and group studies. Such concerns cast doubt on the design-comparability of g_{AB} across SCED and group studies. Maggin et al. (2022) demonstrated several misleading conclusions based exclusively on ES indices. Thus, Maggin et al. (2022) emphasized the importance of conducting systematic visual analyses of single-case data in terms of level, trend, variability, immediacy of change, overlap, and consistency of data patterns both within and across phases.

Given the issues and concerns summarized above, we caution readers not to rely solely on the magnitude of a D-CES, such as g_{AB}, when determining an intervention's effectiveness. Instead, a detailed description of an intervention study should be reviewed thoroughly when assessing an intervention. A study description should provide sufficient information regarding the definition of the construct being intervened, the construct validity of its measurements, the verification of the study's design standards, the examination of the study's operational issues, visual and quantitative demonstrations of the intervention's effectiveness. As Chen et al. (2020), Kratochwill et al. (2021), and Wrigley and McCusker (2019) advocated, which we agree with, that it is more meaningful to ask, "For whom and under what conditions does an intervention work?" than to ask, "Does an intervention work?" A sound reporting of g_{AB}, or any ES index, should include (A) a thorough interpretation of the ES magnitude based on similar studies (i.e., construct and population of interest, treatment/intervention introduced, and measurements of outcomes), and (B) clinical or practical importance of the intervention effect.

The data and materials for the simulation study are available at https://osf.io/hsvwu/.

Appendix A

In this Appendix, we compare our findings under normal distributions to those obtained from Study 2 of Pustejovsky et al. (2014) in terms of relative bias, relative bias of variance estimators, *MSE*, and *CR*.

In terms of g_{AB}'s relative bias, Pustejovsky et al. (2014) reported the average absolute relative bias of g_{AB} to be less than 7.3% when $m = 3$, less than 4.9% when $m = 4$, and less than 2.9% when $m \geq 5$, across all combinations of parameters and N. In its supplementary materials, Pustejovsky et al. (2014) noted that relative bias was generally greater when $N = 8$ than when $N = 16$ (pp. 8–9). Relative bias decreased as m increased. We too observed these patterns in g_{AB}'s relative bias. The average absolute relative biases uncovered in our study were slightly higher than those cited above. The greatest absolute relative bias of g_{AB} was 10.8% when $m = 3$, 7.6% when $m = 4$, 5.5% when $m = 5$, and less than 4.4% when $m = 6$.

In terms of relative bias of g_{AB}'s variance estimator, Pustejovsky et al. (2014) reported substantial overestimation of variance when m was small, averaging 43% when $m = 3$ and $N = 16$. Even at the largest $m = 6$ and $N = 16$, the variance of g_{AB} was overestimated with an average relative bias of 14%. We too observed the persistent overestimation by g_{AB}'s variance estimator. The positive relative biases uncovered in our study were comparable to those cited above. When $m = 3$ and $N = 16$, the average relative bias was 48%. When $m = 6$ and $N = 16$, the average relative bias was 20%.

In terms of g_{AB}'s *MSE* as a criterion for its precision, Pustejovsky et al. (2014, supplementary materials) reported that *MSE* at $N = 8$ ranged from 0.290 ($m = 3$), 0.198 ($m = 4$), 0.149 ($m = 5$), to 0.120 ($m = 6$) across levels of the parameters (p. 9). At $N = 16$, *MSE* ranged from 0.221 ($m = 3$), 0.153 ($m = 4$), 0.116 ($m = 5$), to 0.092 ($m = 6$) also across levels of the parameters. Pustejovsky et al. (2014, supplementary materials) concluded that (1) g_{AB} was imprecise especially when m was small; (2) *MSE*s decreased as m or N increased; and (3) *MSE*s generally increased as ρ, λ, and

[14] Valentine et al. (2016) stated, "In general, when specifying the phase time trends used in the REML model for estimating design-comparable effect sizes, we recommend that users balance prior theory, visual inspection of the data, and parsimony …. we suggest that most users will probably focus on just the two simplest options of no trends or linear trends …. If the user is planning to include the effect size in a synthesis, we recommend that similar model specifications (i.e., phase time trends, and fixed/random effects, as described below) be used for all studies included in the synthesis" (pp. 20-21).

ϕ increased (p. 9). We too uncovered patterns comparable to those cited above. Our *MSE*s at $N=8$ ranged from 0.278 ($m=3$), 0.191 ($m=4$), 0.146 ($m=5$), to 0.119 ($m=6$). At $N=16$, our *MSE*s ranged from 0.222 ($m=3$), 0.156 ($m=4$), 0.120 ($m=5$), to 0.097 ($m=6$).

In terms of *CR*, Pustejovsky et al. (2014) reported that the symmetric CI had greater than nominal coverage, ranging from 97.8% when $m=3$ to 96.4% when $m=6$, averaged across the parameter levels and *N*. When $m=6$, *CR*s varied from 93.9% to 97.6% across levels of the manipulated factors. We too observed the g_{AB}'s CI to over-cover**.** The *CR*s in our study were comparable to those cited above. Our average *CR*s ranged from 97.3% when $m=3$ to 96.6% when $m=6$. When $m=6$, *CR*s ranged from 95.0% to 98.0% across levels of the manipulated factors.

In summary, our results from the normal distribution agreed with Pustejovsky et al.'s (2014), obtained also under the normal distribution, in terms of g_{AB}'s overestimated variance, imprecision when *m* was small, and over-coverage of the symmetric 95% CI. As for g_{AB}'s small relative bias when $m \geq 4$ claimed in Pustejovsky et al. (2014), the average absolute relative bias of g_{AB} in our study fell below 5% only when $m=6$. Subtle differences in findings might be attributed to ways in which Pustejovsky et al. (2014) and this study dealt with non-converged results, and manipulated the autocorrelation (ϕ) slightly differently.

Appendix B

In this appendix, we explain (1) how a case was generated from a normal distribution and (2) how 3 (=*m*) cases were combined to form a replication. We use Pustejovsky et al.'s (2014) matrix notations to facilitate the explanation.

Generating a case

Under MB2, the distribution of the *j*th measurement of the *i*th case (Y_{ij}) is given in Eq. 6 which is repeated here as Eq. B1:

$$\begin{aligned} Y_{ij} &= \gamma_{00} + \eta_{0i} + \left(\gamma_{10} + \eta_{1i}\right) \times D_{ij} + \varepsilon_{ij} \\ &= \left[\gamma_{00} + \left(\gamma_{10} \times D_{ij}\right)\right] + \left\{\eta_{0i} + \left(\eta_{1i} \times D_{ij}\right) + \varepsilon_{ij}\right\} \\ &= [\text{fixed effects}] + \{\text{random effects}\}. \end{aligned} \tag{B1}$$

For *N* scores of a case, Eq. B1 can be expressed in matrix notations as Eq. B2 or 9:

$$\mathbf{Y}_{N\times 1} = \mathbf{D}_{N\times 2}\ \boldsymbol{\gamma}_{2\times 1} + \mathbf{e}_{N\times 1}. \tag{B2}$$

The fixed effects in Eq. B2 are expressed as the product of the design matrix ($\mathbf{D}_{N\times 2}$) and a fixed-effect vector ($\boldsymbol{\gamma}_{2\times 1}$). The random effects in Eq. B2 are expressed as a vector ($\mathbf{e}_{N\times 1}$) that consists of Level-2 errors $[\eta_{0i} + (\eta_{1i} \times D_{ij})]$ and Level-1 errors (ε_{ij}).

For illustration, let's consider the first case from the condition of $m=3$, $N=8$, $\rho=0.2$, $\lambda=0.1$, and $\phi=-0.4$. The **D** and **γ** matrices for this illustrated case are defined according to Eq. B3:

$$\mathbf{D}_{8\times 2} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}, \text{and } \boldsymbol{\gamma}_{2\times 1} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \tag{B3}$$

where the first column of $\mathbf{D}_{8\times 2}$ is a vector of 1s, the second column of $\mathbf{D}_{8\times 2}$ specifies that this case's intervention starts on the fourth measurement (see Table 2), and $\boldsymbol{\gamma}_{2\times 1}$ contains γ_{00} and γ_{10}—the two parameters specified under MB2 (see "Definition of δ_{AB}").

Errors in the error vector ($\mathbf{e}_{8\times 1}$) were randomly generated from a distribution specified in Table 1, using the mvrnonnorm function in the semTools package (Jorgensen et al., 2021). If Dist = normal, random errors were generated from the normal distribution by specifying skewness = 0, kurtosis = 0, and a variance-covariance matrix of errors ($\boldsymbol{\Sigma}_{8\times 8}$ from Eq. B6 below) in the mvrnonnorm function. With these specifications, the mvrnonnorm function produced multivariate normal errors using the Vale and Maurelli method (Vale & Maurelli, 1983). One such $\mathbf{e}_{8\times 1} = (-0.54, 0.30, 0.14, 0.62, -0.07, 1.95, 1.33, 0.87)^{\mathrm{T}}$. Adding this $\mathbf{e}_{8\times 1}$ to the product of $\mathbf{D}_{8\times 2}$ and $\boldsymbol{\gamma}_{2\times 1}$ matrices from Eq. B3, we obtained data for the illustrated case in Eq. B4:

$$\mathbf{Y}_{8\times 1} = \mathbf{D}_{8\times 2}\ \boldsymbol{\gamma}_{2\times 1} + \mathbf{e}_{8\times 1} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} + \begin{bmatrix} -0.54 \\ 0.30 \\ 0.14 \\ 0.62 \\ -0.07 \\ 1.95 \\ 1.33 \\ 0.87 \end{bmatrix} = \begin{bmatrix} -0.54 \\ 0.30 \\ 0.14 \\ 1.62 \\ 0.93 \\ 2.95 \\ 2.33 \\ 1.87 \end{bmatrix}. \tag{B4}$$

The variance-covariance matrix of errors (Σ) is written generally as Eq. B5:

$$\sum = \mathbf{DTD}^{\mathrm{T}} + (1-\rho) \bullet \boldsymbol{AR}(\mathbf{1}), \tag{B5}$$

where $\mathbf{DTD}^{\mathrm{T}}$ = the variance-covariance matrix of Level-2 errors, $(1-\rho) \bullet \boldsymbol{AR}(\mathbf{1})$ = the variance-covariance matrix of Level-1 errors, **D** = design matrix from Eq. B3, $\mathbf{T}_{2\times 2} = \begin{bmatrix} \tau_0^2 & 0 \\ 0 & \tau_1^2 \end{bmatrix}$ (see Footnote 4), and $\boldsymbol{AR}(\mathbf{1})$ is the matrix of

first-order autocorrelations with 1s along the diagonal and $\phi^{|k-j|}$ off-diagonal.

For the illustrated case, $\rho = 0.2$ and $\lambda = 0.1$. Hence, $\tau_0^2 = 0.2$ and $\tau_1^2 = 0.02$ because $(\sigma^2 + \tau_0^2) = 1$ (see "Definition of δ_{AB}"), $\rho = \tau_0^2/(\sigma^2 + \tau_0^2) = \tau_0^2$, and $\lambda = \tau_1^2/\tau_0^2$. With $\phi = -0.4$ and $N = 8$, the $\mathbf{\Sigma}_{8\times8}$ for the eight random errors was computed according to Eq. B6:

$$\mathbf{\Sigma}_{8\times8} = \mathbf{DTD}^{\mathrm{T}} + (1-\rho) \bullet \boldsymbol{AR}(1)$$

$$= \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0.2 & 0 \\ 0 & 0.02 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} +$$

$$(1-0.2) \bullet \begin{bmatrix} 1 & -0.4 & (-0.4)^2 & (-0.4)^3 & (-0.4)^4 & (-0.4)^5 & (-0.4)^6 & (-0.4)^7 \\ -0.4 & 1 & -0.4 & (-0.4)^2 & (-0.4)^3 & (-0.4)^4 & (-0.4)^5 & (-0.4)^6 \\ (-0.4)^2 & -0.4 & 1 & -0.4 & (-0.4)^2 & (-0.4)^3 & (-0.4)^4 & (-0.4)^5 \\ (-0.4)^3 & (-0.4)^2 & -0.4 & 1 & -0.4 & (-0.4)^2 & (-0.4)^3 & (-0.4)^4 \\ (-0.4)^4 & (-0.4)^3 & (-0.4)^2 & -0.4 & 1 & -0.4 & (-0.4)^2 & (-0.4)^3 \\ (-0.4)^5 & (-0.4)^4 & (-0.4)^3 & (-0.4)^2 & -0.4 & 1 & -0.4 & (-0.4)^2 \\ (-0.4)^6 & (-0.4)^5 & (-0.4)^4 & (-0.4)^3 & (-0.4)^2 & -0.4 & 1 & -0.4 \\ (-0.4)^7 & (-0.4)^6 & (-0.4)^5 & (-0.4)^4 & (-0.4)^3 & (-0.4)^2 & -0.4 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & -0.12 & 0.33 & 0.15 & 0.22 & 0.19 & 0.20 & 0.20 \\ -0.12 & 1 & -0.12 & 0.33 & 0.15 & 0.22 & 0.19 & 0.20 \\ 0.33 & -0.12 & 1 & -0.12 & 0.33 & 0.15 & 0.22 & 0.19 \\ 0.15 & 0.33 & -0.12 & 1.02 & -0.10 & 0.35 & 0.17 & 0.24 \\ 0.22 & 0.15 & 0.33 & -0.10 & 1.02 & -0.10 & 0.35 & 0.17 \\ 0.19 & 0.22 & 0.15 & 0.35 & -0.10 & 1.02 & -0.10 & 0.35 \\ 0.20 & 0.19 & 0.22 & 0.17 & 0.35 & -0.10 & 1.02 & -0.10 \\ 0.20 & 0.20 & 0.19 & 0.24 & 0.17 & 0.35 & -0.10 & 1.02 \end{bmatrix} \quad \text{(B6)}$$

The $\mathbf{\Sigma}_{8\times8}$ computed in Eq. B6 was input into the mvrnonnorm function, along with skewness = 0 and kurtosis = 0, to generate normally distributed errors shown in Eq. B4 above.

Generating m cases to form a replication

For a specific *m* in a replication, *m* cases were generated similarly as the illustrated case above. The design matrix $\mathbf{D}_{N\times2}$ of each case varied according to the start point specified in Table 2. After data were generated for *m* cases, a replication was formed and a g_{AB} was computed.

Funding The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of Conflicting Interests The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Algina, J., Keselman, H. J., & Penfield, R. D. (2005). An alternative to Cohen's standardized mean difference effect size: A robust parameter and confidence interval in the two independent groups case. *Psychological Methods, 10*(3), 317–328. https://doi.org/10.1037/1082-989X.10.3.317

American Psychological Association (2020). *JARS—Quant Table 9: Quantitative meta-analysis article reporting standards: Information recommended for inclusion in manuscripts reporting quantitative meta-analyses*. https://apastyle.apa.org/jars/quant-table-9.pdf

Anaby, D., Avery, L., Gorter, J. W., Levin, M. F., Teplicky, R., Turner, L., Cormier, I., & Hanes, J. (2020). Improving body functions through participation in community activities among young people with physical disabilities. *Developmental Medicine & Child Neurology, 62*(5), 640–646. https://doi.org/10.1111/dmcn.14382

Au, T. M., Sauer-Zavala, S., King, M. W., Petrocchi, N., Barlow, D. H., & Litz, B. T. (2017). Compassion-based therapy for trauma-related shame and posttraumatic stress: Initial evaluation using a multiple baseline design. *Behavior Therapy, 48*(2), 207–221. https://doi.org/10.1016/j.beth.2016.11.012

Baek, E., Beretvas, S. N., Van den Noortgate, W., & Ferron, J. M. (2020). Brief research report: Bayesian versus REML estimations with noninformative priors in multilevel single-case data. *The Journal of Experimental Education, 88*(4), 698–710. https://doi.org/10.1080/00220973.2018.1527280

Bandalos, D. L., & Leite, W. (2013). Use of Monte Carlo studies in structural equation modeling research. In G. R. Hancock, & R. D. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 625–666). Information Age Publishing.

Barker, J., McCarthy, P., Jones, M., & Moran, A. (2011). *Single-case research methods in sport and exercise psychology* (1st ed.). Routledge. https://doi.org/10.4324/9780203861882

Barton, E. E., Meadan, H., & Fettig, A. (2019). Comparison of visual analysis, non-overlap methods, and effect sizes in the evaluation of parent implemented functional assessment based interventions. *Research in Developmental Disabilities, 85*, 31–41. https://doi.org/10.1016/j.ridd.2018.11.001

Becraft, J. L., Borrero, J. C., Sun, S., & McKenzie, A. A. (2020). A primer for using multilevel models to meta-analyze single case design data with AB phases. *Journal of Applied Behavior Analysis, 53*(3), 1799–1821. https://doi.org/10.1002/jaba.698

Beretvas, S. N., & Chung, H. (2008). A review of meta-analyses of single-subject experimental designs: Methodological issues and practice. *Evidence-Based Communication Assessment and Intervention, 2*(3), 129–141. https://doi.org/10.1080/17489530802446302

Bolin, J. H., Finch, W. H., & Stenger, R. (2019). Estimation of random coefficient multilevel models in the context of small numbers of level 2 clusters. *Educational and Psychological Measurement, 79*(2), 217–248. https://doi.org/10.1177/0013164418773494

Bollen, K. A., Harden, J. J., Ray, S., & Zavisca, J. (2014). BIC and alternative Bayesian information criteria in the selection of structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(1), 1–19. https://doi.org/10.1080/10705511.2014.856691

Braunstein, S. L. (1992). How large a sample is needed for the maximum likelihood estimator to be approximately Gaussian? *Journal of Physics A: Mathematical and General, 25*, 3813–3826.

Brosnan, J., Moeyaert, M., Newsome, K. B., Healy, O., Heyvaert, M., Onghena, P., & Van den Noortgate, W. (2018). Multilevel analysis of multiple-baseline data evaluating precision teaching as an intervention for improving fluency in foundational

reading skills for at risk readers. *Exceptionality, 26*(3), 137–161. https://doi.org/10.1080/09362835.2016.1238378

Byiers, B. J., Reichle, J., & Symons, F. J. (2012). Single-subject experimental design for evidence-based practice. *American Journal of Speech-Language Pathology, 21*(4), 397–414. https://doi.org/10.1044/1058-0360(2012/11-0036)

Chen, L.-T., 丁麒文, 謝承佑, 陳奕凱, 江宇珊, 黃思婧, 楊同榮, 鄭澈, 劉佩艷, 彭昭英 (2020). 效果量在臺灣心理與教育期刊的應用:回顧與再思[Effect size reporting practices in Taiwanese psychology and education journals: Review and beyond]. 中華心理學刊[*Chinese Journal of Psychology*], *62*(4), 553–592. http://www.cjpsy.com/_i/assets/upload/files/pg066%2B(1).pdf.pdf

Chen, L.-T., Wu, P.-J., & Peng, C.-Y. J. (2019). Accounting for baseline trends in intervention studies: Methods, effect sizes and software. *Cogent Psychology, 6*(1), Article 1679941. https://doi.org/10.1080/23311908.2019.1679941

Chen, M., & Pustejovsky, J. E. (2022). Multilevel meta-analysis of single-case experimental designs using robust variance estimation. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000510

Christoffersen, P. F. (2004). *Elements of financial risk management* (1st ed.). Academic Press. https://doi.org/10.1016/B978-0-12-174232-4.X5000-4

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Cotter, J., & Hanly, J. (2012). Re-evaluating hedging performance for asymmetry: The case of crude oil. In J. Batten & N. F. Wagner (Eds.), *Derivative securities pricing and modeling. Contemporary Studies in Economic and Financial Analysis* (Vol. 94, pp. 259–280). https://doi.org/10.1108/S1569-3759(2012)0000094013

Darbyshire, P., & Hampton, D. (2012). *Hedge fund modelling and analysis using Excel and VBA* (1st ed.). Wiley.

Fan, X., & Fan, X. (2005). Power of latent growth modeling for detecting linear growth: Number of measurements and comparison with other analytic approaches. *The Journal of Experimental Education, 73*(2), 121–139. https://doi.org/10.3200/JEXE.73.2.121-139

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

Franklin, R. D., Allison, D. B., & Gorman, B. S. (1996). Introduction. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 1–12). Lawrence Erlbaum Associates, Inc.

Grasley-Boy, N. M., Gage, N. A., Reichow, B., MacSuga-Gage, A. S., & Lane, H. (2021). A conceptual replication of targeted professional development to increase teachers' behavior-specific praise. *School Psychology Review*. Advance online publication. https://doi.org/10.1080/2372966X.2020.1853486

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

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*(2), 107–128. https://doi.org/10.2307/1164588

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics, 32*(4), 341–370. https://doi.org/10.3102/1076998606298043

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

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research, 26*(3), 329–367. https://doi.org/10.1177/0049124198026003003

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

Jamshidi, L., Heyvaert, M., Declercq, L., Fernández-Castilla, B., Ferron, J. M., Moeyaert, M., Beretvas, S. N., Onghena, P., & Van den Noortgate, W. (2022). A systematic review of single-case experimental design meta-analyses: Characteristics of study designs, data, and analyses. *Evidence-Based Communication Assessment and Intervention*. Advance online publication. https://doi.org/10.1080/17489539.2022.2089334

Jiang, J. (2007). *Linear and generalized linear mixed models and their applications* (1st ed.). Springer. https://doi.org/10.1007/978-0-387-47946-0

Joo, S.-H. (2017). *Robustness of the within-and between-series estimators to non-normal multiple-baseline studies: A Monte Carlo study* (Publication No. 10266637). [Doctoral dissertation, University of South Florida]. ProQuest Dissertations and Theses Global.

Joo, S.-H., & Ferron, J. M. (2019). Application of the within- and between-series estimators to non-normal multiple-baseline data: Maximum likelihood and Bayesian approaches. *Multivariate Behavioral Research, 54*(5), 666–689. https://doi.org/10.1080/00273171.2018.1564877

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2021). *semTools: Useful tools for structural equation modeling*. R package version 0.5-5. Retrieved from https://CRAN.R-project.org/package=semTools

Kazdin, A. E. (2011). *Single-case research designs: Methods for clinical and applied settings* (2nd ed.). Oxford University Press.

Kazdin, A. E. (2019). Single-case experimental designs. Evaluating interventions in research and clinical practice. *Behaviour Research and Therapy, 117*, 3–17. https://doi.org/10.1016/j.brat.2018.11.015

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2013). Single-case intervention research design standards. *Remedial and Special Education, 34*(1), 26–38. https://doi.org/10.1177/0741932512452794

Kratochwill, T. R., Horner, R. H., Levin, J. R., Machalicek, W., Ferron, J., & Johnson, A. (2021). Single-case design standards: An update and proposed upgrades. *Journal of School Psychology, 89*, 91–105. https://doi.org/10.1016/j.jsp.2021.10.006

Kunze, M. G., Machalicek, W., Wei, Q., & St. Joseph, S. (2021). Coaching via telehealth: Caregiver-mediated interventions for young children on the waitlist for an autism diagnosis using single-case design. *Journal of Clinical Medicine, 10*(8), Article 1654. https://doi.org/10.3390/jcm10081654

Lee, J., Bryant, D. P., & Bryant, B. R. (2022). The effect of a Tier 2 multicomponent fraction intervention for fifth graders struggling with fractions. *Remedial and Special Education*. Advance online publication. https://doi.org/10.1177/07419325211069878

Maas, C. J. M., & Hox, J. J. (2004). The influence of violations of assumptions on multilevel parameter estimates and their standard

errors. *Computational Statistics & Data Analysis, 46*(3), 427–440. https://doi.org/10.1016/j.csda.2003.08.006

Maggin, D. M., Barton, E., Reichow, B., Lane, K. L., & Shogren, K. A. (2022). Commentary on the *What Works Clearinghouse Standards and Procedures Handbook* (v. 4.1) for the review of single-case research. *Remedial and Special Education, 43*(6), 421–433. https://doi.org/10.1177/07419325211051317

Man, K., Schumacker, R., Morell, M., & Wang, Y. (2022). Effects of compounded nonnormality of residuals in hierarchical linear modeling. *Educational and Psychological Measurement, 82*(2), 330–355. https://doi.org/10.1177/00131644211010234

McNeish, D. (2017). Small sample methods for multilevel modeling: A colloquial elucidation of REML and the Kenward-Roger correction. *Multivariate Behavioral Research, 52*(5), 661–670. https://doi.org/10.1080/00273171.2017.1344538

Michiels, B., & Onghena, P. (2019). 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., Manolov, R., & Rodabaugh, E. (2020). Meta-analysis of single-case research via multilevel models: Fundamental concepts and methodological considerations. *Behavior Modification, 44*(2), 265–295. https://doi.org/10.1177/0145445518806867

Moeyaert, M., Ugille, M., Ferron, J. M., Beretvas, S. N., & 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., Yang, P., Xu, X., & Kim, E. (2021). Characteristics of moderators in meta-analyses of single-case experimental design studies. *Behavior Modification.* Advanced online publication. https://doi.org/10.1177/01454455211002111

Morgan, D. L., & Morgan, R. K. (2009). *Single-case research methods for the behavioral and health sciences* (1st ed.). SAGE publications.

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine, 38*(11), 2074–2102. https://doi.org/10.1002/sim.8086

Natesan, P. (2019). Fitting Bayesian models for single-case experimental designs: A tutorial. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 15*(4), 147–156. https://doi.org/10.1027/1614-2241/a000180

Natesan, P., & Hedges, L. V. (2017). Bayesian unknown change-point models to investigate immediacy in single case designs. *Psychological Methods, 22*(4), 743–759. https://doi.org/10.1037/met0000134

Onghena, P. (2020). One by one: The design and analysis of replicated randomized single-case experiments. In R. van de Schoot & M. Miočević (Eds.), *Small sample size solutions: A guide for applied researchers and practitioners* (pp. 87–101). Routledge. https://doi.org/10.4324/9780429273872

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

Owens, C. M. & Farmer, J. L. (2013). Analyzing multiple baseline data using multilevel modeling with various residual distributions: A Monte Carlo simulation study. Paper presented at the 2013 annual meeting of American Educational Research Association. Retrieved June 1, 2022, from the AERA Online Paper Repository.

Park, S., & Beretvas, S. N. (2019). Synthesizing effects for multiple outcomes per study using robust variance estimation versus the three-level model. *Behavior Research Methods, 51*(1), 152–171. https://doi.org/10.3758/s13428-018-1156-y

Parker, R. I., & Vannest, K. (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

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling: A Multidisciplinary Journal, 8*(2), 287–312. https://doi.org/10.1207/S15328007SEM0802_7

Peltier, C., Lingo, M. E., Autry, F., Deardorff, M. E., & Palacios, M. (2021). Schema-based instruction implemented under routine conditions. *Journal of Applied School Psychology, 37*(3), 246–267. https://doi.org/10.1080/15377903.2020.1821273

Peltier, C., Lingo, M. E., Deardorff, M. E., Autry, F., & Manwell, C. R. (2020a). Improving word problem solving of immediate, generalized, and combined structured problems via schema-based instruction. *Exceptionality, 28*(2), 92–108. https://doi.org/10.1080/09362835.2020.1727336

Peltier, C., Vannest, K. J., Morin, K. L., Sinclair, T. E., & Sallese, M. R. (2020b). A systematic review of teacher-mediated interventions to improve the mathematical performance of students with emotional and behavioral disorders. *Exceptionality, 28*(2), 121–141. https://doi.org/10.1080/09362835.2020.1771717

Peng, C.-Y. J., Chen, L.-T., Chiang, H.-M., & Chiang, Y.-C. (2013). The impact of APA and AERA guidelines on effect size reporting. *Educational Psychology Review, 25*, 157–209. https://doi.org/10.1007/s10648-013-9218-2

Pustejovsky, J. E., Chen, M., Hamilton, B. J. (2021). *scdhlm: Estimating hierarchical linear models for single-case designs*. University of Wisconsin - Madison, Madison, WI. R package version 0.5.2, https://jepusto.github.io/scdhlm/

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

Pustejovsky, J. E., Swan, D. M., & English, K. W. (2019). An examination of measurement procedures and characteristics of baseline outcome data in single-case research. *Behavior Modification.* Advance online publication. https://doi.org/10.1177/0145445519864264

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical liner models: Applications and data analysis methods* (2nd ed.). SAGE publications.

Rincón, C. L., Muñoz-Martínez, A. M., Hoeflein, B., & Skinta, M. D. (2021). Enhancing interpersonal intimacy in Colombian gay men using functional analytic psychotherapy: An experimental nonconcurrent multiple baseline design. *Cognitive and Behavioral Practice.* Advance online publication. https://doi.org/10.1016/j.cbpra.2021.10.003

Rivera Pérez, J. F., Regalado, A., & Lund, E. (2022). Effects of a computer training to teach Spanish book-sharing strategies to mothers of emergent bilinguals at risk of developmental language disorders: A single-case design study. *American Journal of Speech-Language Pathology, 31*(4), 1771–1786. https://doi.org/10.1044/2022_AJSLP-21-00157

Romano, M. K., & Windsor, K. S. (2020). Increasing deictic gesture use to support the language development of toddlers from high poverty backgrounds. *Early Childhood Research Quarterly, 50*, 129–139. https://doi.org/10.1016/j.ecresq.2018.12.004

Romano, M., Schnurr, M., Barton, E. E., Woods, J., & Weigel, C. (2021). Using peer coaches as community-based competency drivers in Part C early intervention. *Topics in Early Childhood Special Education*. Advance online publication. https://doi.org/10.1177/02711214211007572

Ruiz, F. J., Flórez, C. L., García-Martín, M. B., Monroy-Cifuentes, A., Barreto-Montero, K., García-Beltrán, D. M., Riaño-Hernández, D., Sierra, M. A., Suárez-Falcón, J. C., Cardona-Betancourt, V., & Gil-Luciano, B. (2018). A multiple-baseline evaluation of a

brief acceptance and commitment therapy protocol focused on repetitive negative thinking for moderate emotional disorders. *Journal of Contextual Behavioral Science, 9*, 1–14. https://doi.org/10.1016/j.jcbs.2018.04.004

Saul, J., & Norbury, C. (2021). A randomized case series approach to testing efficacy of interventions for minimally verbal autistic children. *Frontiers in Psychology, 12,* Article 621920. https://doi.org/10.3389/fpsyg.2021.621920

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., & 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

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

Solomon, B. G. (2014). Violations of assumptions in school-based single-case data: Implications for the selection and interpretation of effect sizes. *Behavior Modification, 38*(4), 477–496. https://doi.org/10.1177/0145445513510931

Stewart, N. W., & Hall, C. (2017). The effects of cognitive general imagery training on decision-making abilities in curling: A single-subject multiple baseline approach. *Journal of Applied Sport Psychology, 29*(2), 119–133. https://doi.org/10.1080/10413200.2016.1213331

Tanious, R., & Manolov, R. (2022). A practitioner's guide to conducting and analysing embedded randomized single-case experimental designs. *Neuropsychological Rehabilitation*. Advance online publication. https://doi.org/10.1080/09602011.2022.2035774

Tanious, R., & Onghena, P. (2021). A systematic review of applied single-case research published between 2016 and 2018: Study designs, randomization, data aspects, and data analysis. *Behavior Research Methods, 53*(4), 1371–1384. https://doi.org/10.3758/s13428-020-01502-4

Teh, E. J., Vijayakumar, R., Tan, T. X. J., & Yap, M. J. (2021). Effects of physical exercise interventions on stereotyped motor behaviours in children with ASD: A meta-analysis. *Journal of Autism and Developmental Disorders, 52*, 2934–2957. https://doi.org/10.1007/s10803-021-05152-z

Thurmann-Moe, A. C., Melby-Lervåg, M., & Lervåg, A. (2021). The impact of articulatory consciousness training on reading and spelling literacy in students with severe dyslexia: An experimental single case study. *Annals of Dyslexia, 71*(3), 373–398. https://doi.org/10.1007/s11881-021-00225-1

Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods, 20*(3), 375–393. https://doi.org/10.1037/met0000011

Ugille, M., Moeyaert, M., Beretvas, S. N., Ferron, J., & Van den Noortgate, W. (2012). Multilevel meta-analysis of single-subject experimental designs: A simulation study. *Behavior Research Methods, 44*(4), 1244–1254. https://doi.org/10.3758/s13428-012-0213-1

Ugille, M., Moeyaert, M., Beretvas, S. N., Ferron, J. M., & Van den Noortgate, W. (2014). Bias corrections for standardized effect size estimates used with single-subject experimental designs. *The Journal of Experimental Education, 82*(3), 358–374. https://doi.org/10.1080/00220973.2013.813366

Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika, 48*, 465–471. https://doi.org/10.1007/BF02293687

Valentine, J. C., Tanner- Smith, E. E., Pustejovsky, J. E., & Lau, T. S. (2016). Between-case standardized mean difference effect sizes for single-case designs: A primer and tutorial using the scdhlm web application. *Campbell Systematic Reviews, 12*(1), 1–31. https://doi.org/10.4073/cmdp.2016.1

Verbeke, G., & Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association, 91*(433), 217–221. https://doi.org/10.1080/01621459.1996.10476679

Vlaeyen, J. W. S., Wicksell, R. K., Simons, L. E., Gentili, C., De, T. K., Tate, R. L., Vohra, S., Punja, S., Linton, S. J., Sniehotta, F. F., & Onghena, P. (2020). From boulder to Stockholm in 70 years: Single case experimental designs in clinical research. *The Psychological Record, 70*(4), 659–670. https://doi.org/10.1007/s40732-020-00402-5

What Works Clearinghouse. (2022). *What Works Clearinghouse procedures and standards handbook*, *Version 5.0*. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/WWC-HandbookVer5.0AppIES-508.pdf

Wolfe, K., & McCammon, M. N. (2022). The analysis of single-case research data: Current instructional practices. *Journal of Behavioral Education, 31*(1), 28–42. https://doi.org/10.1007/s10864-020-09403-4

Wrigley, T., & McCusker, S. (2019). Evidence-based teaching: A simple view of "science". *Educational Research and Evaluation, 25*(1-2), 110–126. https://doi.org/10.1080/13803611.2019.1617992

Yuan, K.-H., & Bentler, P. M. (2002). On normal theory based inference for multilevel models with distributional violations. *Psychometrika, 67*(4), 539–561. https://doi.org/10.1007/BF02295130

Zelinsky, N. A. M., & Shadish, W. (2018). A demonstration of how to do a meta-analysis that combines single-case designs with between-groups experiments: The effects of choice making on challenging behaviors performed by people with disabilities. *Developmental Neurorehabilitation, 21*(4), 266–278. https://doi.org/10.3109/17518423.2015.1100690