Optional ERIC Coversheet — Only for Use with U.S. Department of Education Grantee Submissions

This coversheet should be completed by grantees and added to the PDF of your submission if the information required in this form **is not included on the PDF to be submitted**.

INSTRUCTIONS

- Before beginning submission process, download this PDF coversheet if you will need to provide information not on the PDF.
- Fill in all fields—information in this form **must match** the information on the submitted PDF and add missing information.
- Attach completed coversheet to the PDF you will upload to ERIC [use Adobe Acrobat or other program to combine PDF files]—do not upload the coversheet as a separate document.
- Begin completing submission form at https://eric.ed.gov/submit/ and upload the full-text PDF with attached coversheet when indicated. Your full-text PDF will display in ERIC after the 12-month embargo period.

GRANTEE SUBMISSION REQUIRED FIELDS

Title of article, paper, or other content

Between-Case Standardized Mean Differences: Flexible Methods for Single-Case Designs

All author name(s) and affiliations on PDF. If more than 6 names, ERIC will complete the list from the submitted PDF.

Last Name, First Name	Academic/Organizational Affiliation	ORCID ID
Chen, Man	University of Wisconsin - Madison	
Pustejovsky, James	University of Wisconsin - Madison	
Klingbeil, David	University of Wisconsin - Madison	
Van Norman, Ethan	Lehigh University	

Publication/Completion Date—(if *In Press,* enter year accepted or completed) 2023

Check type of content being submitted and complete one of the following in the box below:

- ☑ If article: Name of journal, volume, and issue number if available
- ☐ If paper: Name of conference, date of conference, and place of conference
- ☐ If book chapter: Title of book, page range, publisher name and location
- ☐ If book: Publisher name and location
- ☐ If dissertation: Name of institution, type of degree, and department granting degree

Journal of School Psychology, vol. 98

DOI or URL to published work (if available) https://doi.org/10.1016/j.jsp.2023.02.002

Acknowledgement of Funding— Grantees should check with their grant officer for the preferred wording to acknowledge funding. If the grant officer does not have a preference, grantees can use this suggested wording (adjust wording if multiple grants are to be acknowledged). Fill in Department of Education funding office, grant number, and name of grant recipient institution or organization.

"This work was supported by U.S. Department of Education **[Office name]** Institute of Education Sciences through **[Grant number]** R305D190023 to **Institution]** Lehigh University. The opinions expressed are those of the authors and do not represent views of the **[Office name]** Institute of Education Sciences or the U.S. Department of Education.

Between-Case Standardized Mean Differences: Flexible Methods for Single-Case Designs

Man Chen[1], James E. Pustejovsky[1], David A. Klingbeil[1], and & Ethan R. Van Norman[2]

[1] University of Wisconsin-Madison

[2] Lehigh University

Author Note

Man Chen, Department of Educational Psychology, University of Wisconsin-Madison. James E. Pustejovsky, Department of Educational Psychology, University of Wisconsin-Madison. David A. Klingbeil, Department of Educational Psychology, University of Wisconsin-Madison. Ethan R. Van Norman, Department of Education and Human Services, Lehigh University.

Funding: This work was supported, in part, by the Institute of Educational Sciences, U.S. Department of Education through grant R305D190023 to Lehigh University. The opinions expressed are those of the authors and do not represent the views of the Institute of the U.S. Department of Education.

Declaration of Interests: None.

Correspondence concerning this article should be addressed to James E. Pustejovsky, 1082C Educational Sciences, 1025 W Johnson St. Madison, WI 53706-1706. E-mail: pustejovsky@wisc.edu

Abstract

Single-case designs (SCDs) are a class of research methods for evaluating the effects of academic and behavioral interventions in educational and clinical settings. Although visual analysis is typically the first and main method for primary analysis of data from SCDs, quantitative methods are useful for synthesizing results and drawing systematic generalizations across bodies of single-case research. Researchers who are interested in synthesizing findings across SCDs and between-group designs might consider using the between-case standardized mean difference (BC-SMD) effect size, which aims to put results from both types of studies into a common metric. Currently available BC-SMD methods are limited to treatment reversal designs with replication across participants and across-participant multiple baseline designs, yet more complex designs are used in practice. In this study, we extend available BC-SMD methods to several variations of the multiple baseline design, including the replicated multiple baseline across behaviors or settings, the clustered multiple baseline design, and the multivariate multiple baseline across participants. For each variation, we describe methods for estimating BC-SMD effect sizes and illustrate our proposed approach by re-analyzing data from a published SCD study.

Keywords: single-case design; effect size; between-case standardized mean difference; multiple baseline design

Between-Case Standardized Mean Differences: Flexible Methods for Single-Case Designs

Meta-Analysis of Single-Case Designs

Single-case designs (SCDs) are a class of research methods for evaluating the effects of academic and behavioral interventions in educational and clinical settings (Horner et al., 2005). Researchers using an SCD measure outcomes repeatedly over time, within and across distinct treatment conditions or phases (such as baseline and intervention phases), for each of several cases (Gast & Ledford, 2018). Inferences about treatment effects for each case are drawn by comparing the data series of outcomes in the baseline phase(s) with the pattern in the treatment phase(s), with each case serving as its own control. In other words, SCDs are designed to assess whether there is a functional relation between the implementation of a treatment and a change in the targeted outcome.

SCDs are important and frequently used within certain areas of education and psychology, including applied behavior analysis, school psychology, and early childhood special education (Barton et al., 2016; Kratochwill & Levin, 2014). The designs are attractive in such settings because they can be applied with a small number of participants—even with only a single participant—and because they allow for adaptation, modification, and individualization. Due to their flexibility, SCDs are particularly useful for investigating interventions for individuals with low incidence disabilities. Because SCDs provide a rigorous method for evaluating the causal impact of interventions, they are well-suited for enhancing our knowledge about what works, for whom, and under what conditions (Skinner et al., 2013).

Multiple baseline designs are the most common type of SCD (Barker et al., 2013; Shadish & Sullivan, 2011). Multiple baseline designs involve collecting data across several cases, which may represent different individuals, groups of individuals, or a single individual observed in different settings or for different behaviors. For example, in a

multiple baseline design across participants, baseline data are collected for several participants, and treatment is introduced to one participant after observing a stable baseline phase. Treatment is introduced to another participant once a change in responding is observed for the first participant, and this process is repeated across all participants. The defining feature of all multiple baseline designs is that treatment is introduced at several different time points for different cases. This staggered introduction ensures that treatment is not confounded by contemporaneous changes in the environment, thus protecting against common history threats (Gast et al., 2018).

Visual analysis is typically the first and main method for primary analysis of data from SCDs (Kratochwill et al., 2010, 2021). According to commonly used guidelines, visual analysis of SCD data should investigate six features of the data: level, trend, variability within phases, immediacy of the treatment effect, overlap in data series, and consistency of data patterns between or across phases (Horner et al., 2005; Kratochwill et al., 2010). If those features are consistent with hypothesized changes in the outcome, then a functional relation is established between the treatment and the outcome. Although visual analysis is viewed as a necessary first step in determining whether functional relations are present in primary analysis of individual studies, it does not provide a clear means for quantifying the magnitude of treatment effects or for synthesizing findings from multiple studies that investigate conceptually similar interventions. Consequently, there is interest in using quantitative methods to synthesize results and draw systematic generalizations across bodies of single-case research.

Meta-analyses have been used to synthesize findings across SCD studies in order to provide a firmer basis for generalization about effects of intervention, thus informing identification of evidence-based practices for policy and decision-making (Pustejovsky & Ferron, 2017). Meta-analyses involve summarizing findings from individual studies using quantitative indices, called effect size measures, that describe the magnitude and direction of intervention effects. In addition to providing summaries of findings across studies,

meta-analyses can be used to characterize the extent of variation in effect sizes and to identify systematic predictors of effectiveness.

A variety of effect size measures have been proposed for SCDs (Moeyaert et al., 2018; Pustejovsky & Ferron, 2017). For example, in SCD studies published in school psychology journals, the most frequently reported effect size indices are non-overlap measures (Radley et al., 2020), including the percentage of non-overlapping data (PND, Scruggs et al., 1987), percentage of all non-overlapping data (PAND, Parker et al., 2007), non-overlap of all pairs (NAP, Parker & Vannest, 2009), and baseline-corrected Tau (Tarlow, 2017). These non-overlap statistics summarize intervention effects for each case by measuring the degree of non-overlap between baseline and intervention phases. However, some commonly used non-overlap measures such as PND and PAND are sensitive to incidental procedural characteristics of the study's design. Some also lack known sampling distributions, which limits their utility for meta-analysis of SCDs (Pustejovsky, 2019). Other common indices include the within-case standardized mean difference (WC-SMD, Busk & Serlin, 1992), which describes changes in treatment relative to the within-case variability in the outcome, and the log response ratio (Pustejovsky, 2018), which describes treatment effects in terms of proportional change from baseline. All of these indices are case-specific measures that provide a quantitative summary of the intervention effects for each individual case within a single-case study. Thus, they are on metrics specific to SCDs and not directly comparable to effect sizes from between-group experimental designs.

Between-case standardized mean difference

The standardized mean difference (SMD), often referred to as Cohen's d, is the most widely used effect size measure for quantifying treatment effects in between-group designs. It can be defined as the difference between the population mean outcome if every participant received treatment and the population mean outcome if nobody received treatment, standardized by the standard deviation of the outcome if nobody received

treatment (Hedges, 2008). In the context of simple between-group experimental designs, the SMD is often estimated using the difference in the mean outcomes between the experimental group and the control group, scaled by the square root of the pooled sample variance of the outcome measure, although other estimators are also used for more complex designs (Taylor et al., 2021). For instance, Wright et al. (2012) conducted a randomized controlled trial to examine the effects of video-based self-evaluation package on use of general or specific praise among Head Start teachers. The authors randomly assigned 51 Head Start teachers to an immediate self-evaluation, delayed self-evaluation, or control group. Teachers in the immediate or delayed self-evaluation experimental groups received training in identifying praise, evaluated their videotaped performance, and wrote goals for improving use of praise, while teachers in the control group did not receive such training or self-evaluation package. Although not reported in the article, SMD effect sizes can be calculated by taking the difference in the average frequency of praise statements between the experimental groups and the control group after the intervention was implemented for 3 days. The SMD effect size estimates would describe the effects at Day 3 of implementing the self-evaluation package.

The between-case standardized mean difference (BC-SMD) for SCDs is an effect size metric that is theoretically comparable to a standardized mean difference from a between-group design performed with the same population, the same intervention, and the same outcome measures. For example, Grasley-Boy et al. (2021) used a multiple baseline across participants to examine the effects of targeted professional development and performance feedback on use of behavior-specific praise among elementary school teachers. Considering its similarity with Wright et al. (2012), the researchers here could, in principle, have conducted a between-group experimental design to study the effects of this intervention by 1) randomizing participating teachers to either receive the intervention or continue with their usual practice and 2) assessing frequency of behavior-specific praise for each participating teacher after some period of time (cf. Simonsen et al., 2020; Wright et

al., 2012). Grasley-Boy et al. (2021) reported a BC-SMD of 1.93 (SE = 0.38) standard deviations (Grasley-Boy et al., 2021, p. 9), which can be interpreted in the same way as the SMD from a between-group study on the same topic.

The BC-SMD effect size was initially developed for reversal designs with replication across participants (Hedges et al., 2012) and multiple baseline designs across participants (Hedges et al., 2013). These initial developments were limited by some restrictive assumptions, including that baseline data are stable without trends, that trends do not exist during the treatment phase(s), and that treatment effects do not vary across cases. To resolve these limitations, Pustejovsky et al. (2014) introduced a more general framework based on hierarchical linear models that allow linear or polynomial time trends and heterogeneous treatment effects across cases. Under this framework, SCD data are modeled at both the within-case and between-case levels, and a BC-SMD effect size is calculated as the mean difference between outcomes from different treatment conditions, standardized by the square root of the sum of within- and between-case variance components. Standardizing in this way puts the mean difference on the same scale as the SMD from a between-group experiment.

Because of its theoretical comparability, the BC-SMD effect size provides a way to describe intervention effects in terms more familiar to researchers who predominantly use group designs (Shadish et al., 2015; Shadish, 2014). For interventions that are evaluated using both single-case and group designs, the BC-SMD might also provide a means to compare intervention effects across design types. Thus, use of BC-SMDs can enrich research syntheses and evidence-based practice reviews by allowing for inclusion of evidence from SCDs that might otherwise be omitted (Shadish et al., 2015). Due to these advantages, BC-SMD effect sizes have been applied in several recent meta-analyses of single-case research (e.g., Babb et al., 2021; Losinski et al., 2017; Shin et al., 2020). Furthermore, the What Works Clearinghouse recently adopted the BC-SMD as a metric for summarizing findings from SCD studies (What Works Clearinghouse, 2020). The What

Works Clearinghouse's adoption of BC-SMDs is controversial, with scholars criticizing the approach for de-emphasizing visual analysis and limiting eligibility to SCDs for which BC-SMDs can be calculated (Kratochwill et al., 2021; Maggin et al., 2021). To date, BC-SMD methods have only been formally developed for two of the numerous types of SCDs that are used to investigate functional relations. Developing methods for other variations of SCDs would help to alleviate this problem, both for the What Works Clearinghouse's work and for the broader field.

Novel extensions for the BC-SMD effect size

The BC-SMD methods described in Pustejovsky et al. (2014) apply to multiple baseline designs across participants. In this paper, we will show that the same general framework can also be applied to several more complex forms of SCDs, including the replicated multiple baseline across behaviors or settings, the clustered multiple baseline, and the multivariate multiple baseline design across participants We recognize that these designs are less common than the multiple-baseline across participants design. Yet, they offer an approach to increase the methodological rigor or are well-matched to how interventions are conducted in schools and other applied settings. Extending the BC-SMD to these designs may therefore assist school psychologists and other researchers in the effort to document what works, for whom, under what conditions.

Our first extension is for the replicated multiple baseline design across behaviors or settings, where a multiple baseline design is conducted using cases comprised of different behaviors or settings from a single participant, and the entire design is replicated across multiple participants. This type of design is more rigorous than the multiple-baseline across participants design because it affords both intra-participant and between-participant replication (Gast et al., 2018). As an example, Thiemann and Goldstein (2001) investigated the effects of written text and pictoral cues supplemented by video feedback on the social communication of five students with autism. For each student, baseline data

were collected on several behaviors and intervention was introduced across those behaviors at staggered time points. This process was replicated across five students. As we demonstrate, these data can be modeled with a three-level hierarchical structure with measurements at the first level, behaviors at the second level, and students at the third level. In this design, replication across participants is required for establishing the theoretical comparability between SCDs and between-group designs.

Our second extension is for the clustered multiple baseline design, in which each case is comprised of multiple individuals, such as a small group, classroom, or school, and where outcomes are measured on individuals within each group. Many school-based interventions targeting academic skills (e.g., reading comprehension) or social-behavioral outcomes are implemented in small group formats, making this design useful for examining academic interventions under naturalistic conditions. For example, Bryant et al. (2018) evaluated the effects of an intensive mathematics intervention for second grade students with severe mathematics difficulties. The study included 33 second grade students from 12 groups in 5 schools. The intervention was implemented at the group level. In this example, the data can be modeled with a three-level hierarchical structure with measurements at the first level, participants at second level, and groups at the third level. For purposes of quantifying treatment effects, explicitly modeling this type of clustered structure (rather than ignoring the clustering structure or aggregating the data to the group level) is more consistent with how the intervention was implemented.

Our third extension is for the multivariate multiple baseline design across participants, where treatment effects are investigated in an across-participant design for several distinct outcome measures. This type of design is useful when studying the effects of multi-component interventions that target related academic skills (e.g., reading rate and reading accuracy) or social-behavioral interventions targeting both engagement and task completion. For example, Calder and colleagues (2020) used such a design to investigate the effects of an explicit grammar intervention for nine children with developmental

language disorder across several different outcomes, including expressive morphosyntax production and grammaticality judgements. This data structure can be described as having multivariate outcome measures at the first level, with participants at the second level. The data can be modeled by allowing for correlation among outcome measures collected at each time point, as well as for heterogeneous variances across outcome measures. Extending the BC-SMD to this design would allow researchers to estimate the effect of an intervention on each dependent variable in a single model rather than running multiple, separate analyses, while ignoring the correlation between outcome measures.

Because these three variants of the multiple baseline design are especially relevant for SCDs conducted in school settings, it would be useful to have methods for effect size estimation that can be applied to such designs. Thus, the purpose of the present study is to extend the framework of Pustejovsky et al. (2014) and demonstrate how to calculate BC-SMD effect sizes for these three variations of the multiple baseline design. For each variation, we propose a modeling approach, define a BC-SMD effect size, and develop methods for estimating the BC-SMD. We illustrate our proposed approach to each design by re-analyzing data from a published SCD study.

The current study contributes to the literature in at least two respects. First, although Pustejovsky et al. (2014) proposed a general framework for defining and estimating BC-SMD effect sizes based on hierarchical linear modeling, no methodological research has examined how to apply this framework to more complex SCDs, such as the three variations of the multiple baseline design that we have described. We provide an initial demonstration of how to apply the general methods to each variation, thus contributing to the further development of design-comparable effect sizes. Second, these novel demonstrations provide a template for calculating BC-SMD effect sizes in more complex designs, making it possible to include them in meta-analyses and thereby enriching evidence-based practice reviews.

In the sections that follow, we first review the general BC-SMD framework described in Pustejovsky et al. (2014). We then demonstrate how to extend the existing BC-SMD methods to handle the three variations of multiple baseline designs in terms of model specification and effect size calculation. For each variation, we demonstrate the application of the methods using data from a published SCD study. In the final section, we discuss implications, limitations, and directions for future research.

General Methods

Pustejovsky et al. (2014) proposed methods for estimating BC-SMD effect sizes for data from across-participant multiple baseline and multiple probe designs. The methods entail first estimating a hierarchical linear model (HLM) on the data, then using the estimated model parameters to calculate an effect size estimate. A design with multiple participants is required in order to be able to estimate the degree of between-participant variation in the outcome, which is needed for estimating an effect size in the same metric as the SMD from a between-group design. Pustejovsky et al. (2014) conceptualized and demonstrated the approach based on a two-level hierarchical linear model, which allows modeling time trends within each phase as well as variation in the level, trends, and treatment effects across participants. However, this same effect size estimation strategy can be applied to hierarchical models with more complex structure, such as three-level models or multivariate models, so long as they still include between-participant variation in the outcome. Before illustrating how to extend the methods to more complex SCDs, we first review the framework for specifying the BC-SMD effect size and the general estimation strategy used by Pustejovsky et al. (2014).

In general terms, the BC-SMD effect size is defined by first identifying an appropriate model for the single-case design data, then using that model to consider a hypothetical between-group design conducted with the same population of participants, the same intervention, and the same (or similar) outcome measurement procedures. If the

model for the single-case design is well-specified, then one can use it to estimate the effect size that would be identified in the hypothetical between-group study.

For the first step, we focus on using hierarchical linear models because of the hierarchical structure of SCD data, where outcome measurements are nested within each case. Consider a multiple baseline design across participants. At the first (within-participant) level of the HLM, outcomes for each participant can be modeled as a function of an intercept and a treatment indicator. The intercept represents the average outcomes in the baseline phase and the slope for the treatment indicator represents the treatment effect for each participant. This first-level model specification can be extended in many ways, such as by including an intercept, a treatment indicator, a time variable, and an interaction between treatment indicator and time. In such a model, the treatment effect for a participant is then a combination of the immediate change in the outcome measures and the additional change across time. At the second (between-participant) level of the model, one or more of the first level coefficients can vary across the higher level units. For example, the average outcomes in the baseline phase might vary from participant to participant. The treatment effects might also differ for different participants or the changes in the outcome across measurement occasions might vary across participants. As we demonstrate in subsequent sections, SCD data with more complex structure can be modeled by replacing the two-level HLM with a more flexible model that captures not only how outcomes change as a function of time and treatment phase but also how these relations vary across higher level units.

Model specification

In general, HLMs (including two-level models and more complex models) have two distinct sets of parameters: fixed effect parameters, which describe the average relation between the outcome and a set of predictors, and random effects variance components, which describe how the outcomes vary around their predicted levels (see Snijders & Bosker,

2011, Chapters 6, p.94–108). Specification of an HLM entails making choices about both components of the model. In the context of models for multiple baseline designs, one needs to decide whether to include baseline time trends, change in the time trends from baseline to treatment phase, along with the treatment indicator, in the fixed part at the first level of the model. As for the random part, one needs to consider whether the lower-level parameters vary across higher level units. Considering that various models might be specified for SCD data and that different specifications would lead to different effect size estimates, we suggest researchers use theory and disciplinary knowledge to guide the model specification. Additionally, we recommend researchers also use visual analysis for identifying potential models and conduct preliminary analyses for model comparison (Baek et al., 2016; Moeyaert et al., 2020). Such preliminary analyses might include visual inspection of the model's predicted values, likelihood ratio tests comparing nested models, or examination of model fit statistics such as the Akaike information criterion (AIC) or Bayesian information criterion (BIC).

Effect size parameter

Having arrived at an appropriate model, the remaining question is how to define the hypothetical between-group study. To fully operationalize this, we will need to specify when treatment would first be initiated and when follow-up outcome assessment would occur in the hypothetical study. Let A denote the time point immediately before the start of treatment, and let B denote the time point at which outcomes would be measured in this hypothetical study. The SMD effect size then represents the effect at time B of introducing treatment after time A. In multiple baseline designs, the BC-SMD describes the same effect and is calculated by taking the difference between the model's predictions of the average outcome at time B if treatment were to be introduced after time A and the average outcome at time B if treatment were never introduced. This difference is scaled by the square root of the total variance in the outcomes at time B if treatment were never

introduced (Pustejovsky et al., 2014). Although an actual multiple baseline study involves staggered intervention start points and multiple repeated measurements, all of the required quantities still correspond to components of the model for the multiple baseline data.

Given an appropriate model and a hypothetical between-group design, we can use HLM model parameters to define the BC-SMD effect size parameter. Pustejovsky et al. (2014) showed that a BC-SMD effect size can be defined in terms of the fixed effect parameters and variance component parameters of the model. Letting $\gamma_1, \gamma_2, ..., \gamma_P$ denote the P fixed effect parameters and $\theta_1, \theta_2, ..., \theta_R$ denote the set of R variance components, the BC-SMD effect size parameter δ_{AB} can be represented as

$$\delta_{AB} = \frac{\sum_{a=1}^{P} p_a \gamma_a}{\sqrt{\sum_{b=1}^{R} r_b \theta_b}}. \tag{1}$$

In this expression, $p_1, ..., p_P$ are constants that determine how to weight each of the fixed effects parameters to calculate the effect size. Likewise, $r_1, .., r_R$ are constants that determine whether to use each of the variance components—including the random effects variances and covariances, the variance structure parameters, and the level-1 error variances—for calculating the effect size. These constants depend on the hypothetical experimental features, A and B, as well as on the form of the data model.

Selecting A and B

The selection of times A and B can influence the magnitude of BC-SMD effect size estimates when baseline time trends vary across participants or when time trends differ in magnitude between the baseline and intervention phases (Valentine et al., 2016). In a between-group design, the study intake time will typically be somewhat arbitrary, and so the same holds for the choice of initial starting time A for the hypothetical design under consideration. The default choice for A in the *scdhlm* web application (Pustejovsky et al., 2020a) is the last measurement occasion of the shortest baseline phase across cases, and we recommend following this convention unless one has an explicit justification for doing

otherwise. We also note that this choice of time point is at least broadly consistent with the logic of SCD, which emphasizes that cases should have stable baseline phases before introducing the intervention phase (Gast et al., 2018).

Follow-up time B should ideally be chosen by following the precedents of previous research that uses similar interventions for the same population–either from between-group designs or SCDs. In the context of meta-analyses, researchers can also examine the typical intervention durations used in the included studies and choose a follow-up time based on a common benchmark duration. If no convention or guideline is available, researchers might use the default choice of B in the *scdhlm* web application (Pustejovsky et al., 2020a), which corresponds to the shortest treatment duration across cases. This default is designed to limit the extent of extrapolation. In models where the choice of time-points matters, researchers can and should conduct sensitivity analyses to further investigate the extent to which BC-SMD effect size estimates are influenced by time A and B. We comment further on the choice of time A and B in the Discussion.

Estimation

Estimation of the BC-SMD effect size entails first estimating the parameters of the model, along with their sampling variances, then substituting the estimates for the corresponding model parameters, and finally making a small-sample correction (Pustejovsky et al., 2014). Following typical practices for estimating HLMs, the unknown variance component parameters $(\theta_1, ..., \theta_R)$ can be estimated using restricted maximum likelihood (REML) techniques. The sampling variance-covariance matrix of the variance parameters can be approximated with the inverse of the expected Fisher information matrix based on REML estimators (Gilmour et al., 1995). The fixed effect coefficients $(\gamma_1, ..., \gamma_P)$ are then estimated using weighted least squares, treating the variance component estimates as if they were known. The sampling variance-covariance matrix of the fixed effect coefficients can be approximated based on the variance component estimates.

The BC-SMD effect size in Equation (1) is estimated by replacing the fixed effect parameters and the variance component parameters with the corresponding estimators from the fitted model. However, this effect size estimator may be biased if it is based on a small number of cases, as typically found in SCDs. Pustejovsky et al. (2014) proposed a small-sample correction to obtain a less-biased estimator, which involves approximating the sampling distribution of the estimator by a t distribution with degrees of freedom given by a Satterthwaite approximation. Let $\hat{\gamma}_1, ..., \hat{\gamma}_P$ and $\hat{\theta}_1, ..., \hat{\theta}_R$ denote the REML estimators of the fixed effects and variance components, respectively. Let $\mathrm{C}(\hat{\gamma}_a, \hat{\gamma}_b)$ denote the estimated sampling covariance between fixed effect estimates $\hat{\gamma}_a$ and $\hat{\gamma}_b$; let $\mathrm{C}(\hat{\theta}_a, \hat{\theta}_b)$ denote the estimated sampling covariance between the variance component estimates $\hat{\theta}_a$ and $\hat{\theta}_b$. The bias-corrected effect size estimator g_{AB} is given by

$$g_{AB} = J(\nu) \times \frac{\sum_{a=1}^{P} p_a \hat{\gamma}_a}{\sqrt{\sum_{b=1}^{R} r_b \hat{\theta}_b}}, \tag{2}$$

where $J(\nu) = \left(1 - \frac{3}{4\nu - 1}\right)$ and ν is the Satterthwaite degrees of freedom given by

$$\nu = \frac{2\left(\sum_{a=1}^{R} r_a \hat{\theta}_a\right)^2}{\sum_{a=1}^{R}\sum_{b=1}^{R} r_a r_b \mathrm{C}(\hat{\theta}_a, \hat{\theta}_b)}$$

(Pustejovsky et al., 2014). An approximate standard error for the effect size estimator g_{AB} is

$$\mathrm{SE}_g = J(\nu)\sqrt{\frac{\nu\kappa^2}{\nu - 2} + g_{AB}^2\left(\frac{\nu}{\nu - 2} - \frac{1}{J(\nu)^2}\right)}, \tag{3}$$

where

$$\kappa = \sqrt{\frac{\sum_{a=1}^{P}\sum_{b=1}^{P} p_a p_b \mathrm{C}(\hat{\gamma}_a, \hat{\gamma}_b)}{\sum_{b=1}^{R} r_b \hat{\theta}_b}}.$$

A $(1-\alpha)$ symmetric or asymmetric confidence interval (CI) for the effect size estimator g_{AB} can be constructed using a central or noncentral t approximation, respectively (see Pustejovsky et al., 2014 for details).

These general estimation methods can be applied to estimate BC-SMDs from designs more complex than the multiple baseline designs across participants. To do so, we will first need to identify appropriate models for each design, then determine how the fixed effects and variance component parameters of those designs should be combined in calculating a BC-SMD. This second step is essentially the question of what values to use for the constants $p_1, ..., p_P$ and $r_1, ..., r_R$. In the following sections, we demonstrate these steps for several types of multiple baseline designs that have more complex features.

Replicated Multiple Baseline Across Behaviors or Settings

One common type of multiple baseline design is the multiple baseline across behaviors or settings. In this design, cases correspond to different behaviors or behavior sets (or to the same behavior in different settings) of a single participant or group of participants (Gast et al., 2018). Baseline data are collected on each of the behaviors and intervention is introduced for each behavior at a different point in time. Although this design can be conducted with only a single participant, it is common to report replications of the same design across several participants. In a replicated multiple baseline across behaviors or settings, we can calculate a BC-SMD that reflects the average intervention effect across behaviors and across participants. Just as with other types of designs, replication across participants is necessary here so that between-participant variation in the outcome can be estimated and comparability with the effect size from a between-group

design can be achieved.

Thiemann and Goldstein (2001) used a replicated multiple baseline across behaviors to examine the effectiveness of visually cued instruction on social communication of five children with autism and social deficits. For each participant, treatment was implemented at a different time for each of four social communication behaviors: contingent responses, securing attention, initiating comments, and initiating requests. Two behaviors, initiating comments and initiating requests, were combined and examined as one behavior set for Casey and John due to low initiations in the baseline phase. The outcome variable was frequency of occurrence of these social communication behaviors over a 10 minute social activity, coded using a 15-s whole interval system. Our analysis excludes the data of two behaviors for two participants, which were collected in the baseline phase only. Data for the five participants are displayed in Figure 1, with the data for each participant depicted using a distinct color.

Model specification

In order to estimate a BC-SMD effect size from a replicated multiple baseline across behaviors, we must first identify a model for the observed data. Generally, data from these designs can be described by a three-level hierarchical model, with outcome measurements at level 1, behaviors at level 2, and participants at level 3. For purposes of illustration, we will discuss effect size estimation using a relatively simple specification that still captures the primary features of the data from Thiemann and Goldstein (2001).

We used visual inspection of the data to inform our model specification. For the within-participant level (i.e., level 1) of the model, we need to consider whether to include time trends in the baseline phase or intervention phase. Figure 1 shows that both baseline and treatment phase are relatively stable for most behaviors of most participants, especially for Casey, Greg, and Ivan. Thus, we will tentatively assume that there are no systematic time trends in the baseline phases nor changes in time trends from baseline to treatment

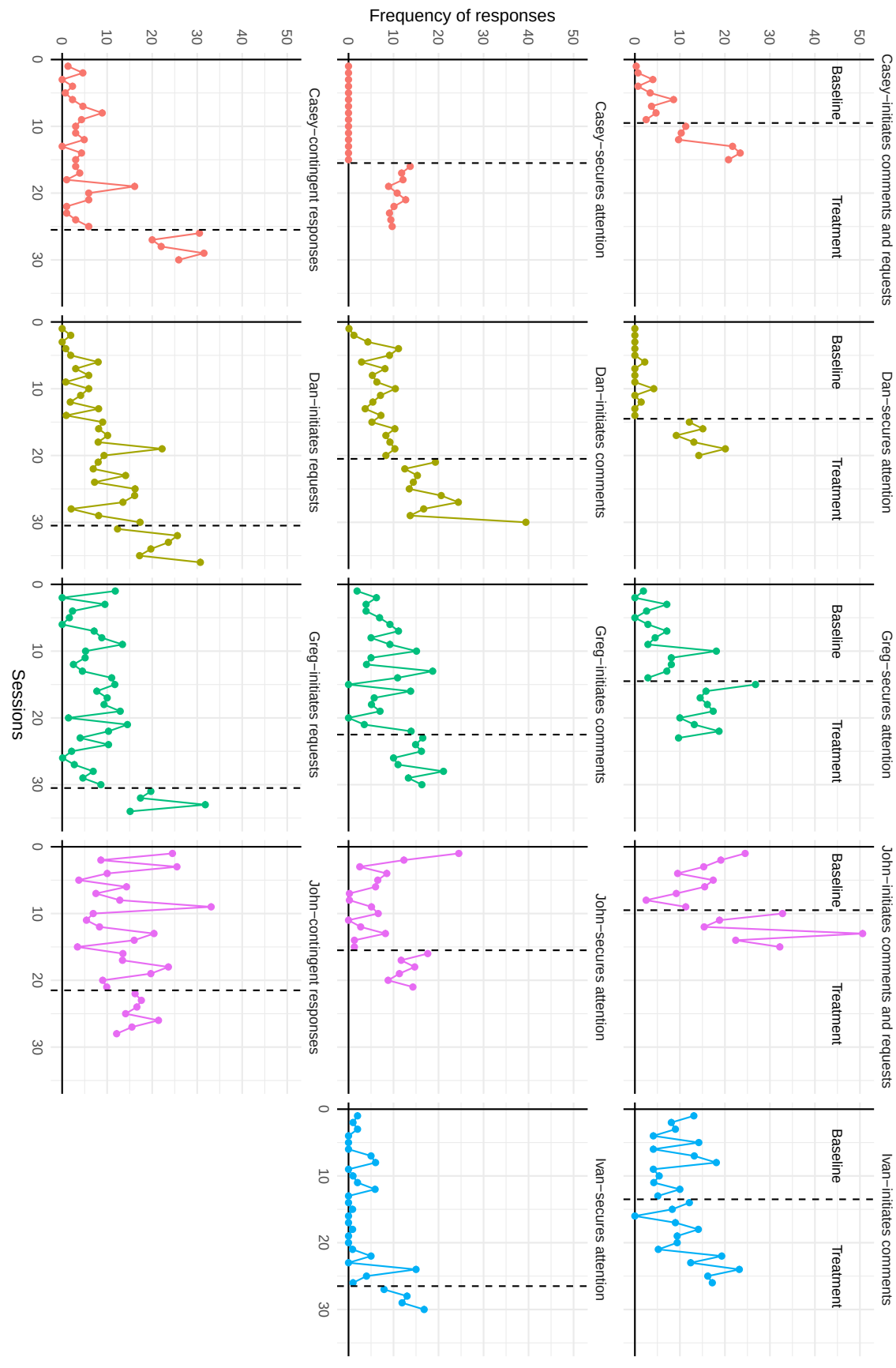

Figure 1

Multiple baseline design across behaviors from Thiemann and Goldstein (2001).

phase. For the between-participant level (i.e., level 2), we need to consider whether the level 1 parameters vary across cases. Visual inspection suggests that both the average outcomes in the baseline phase and the changes in the average outcomes in the treatment phase seem to vary across behaviors within participants and across all five participants. Therefore, we use a model that includes random effects for both of these components.

Let Y_{tsi} denote the outcome measure at time t for behavior, setting, or series s for participant i, where $t = 1, \ldots, N$, $s = 1, \ldots, S_i$, and $i = 1, \ldots, m$. Let T_{si} denote the time-point after which treatment is introduced for series s for participant i, and let $I(t > T_{si})$ be a dummy variable equal to 0 during the baseline phase and to 1 during the treatment phase. Based on our preliminary visual inspection of the data, we specify a model without time trends. The model for the outcome measurements in series s of participant i can therefore be expressed as:

$$Y_{tsi} = \pi_{0si} + \pi_{1si} I(t > T_{si}) + e_{tsi}, \tag{4}$$

where π_{0si} is the average level of the outcome for series s of case i in the absence of treatment and π_{1si} is the change in the outcome measure in series s for case i upon introduction of treatment. Because outcomes are measured repeatedly over time for each case, it might be considered implausible that the errors are independent. Following the convention for analysis of SCD data, we assume that the error term e_{tsi} has mean zero, variance σ^2, and first-order auto-correlation ϕ (i.e., a first-order auto-regression process).

To complete the model specification, we need to determine how the parameters of Equation (4) vary across series and across participants. Based on visual analysis, we allow both π_{0si} and π_{1si} to vary across series. That is, the model allows for random variation across series in the intercepts and treatment effects of each participant. This can be

expressed as the level-2 model

$$\pi_{0si} = \beta_{00i} + r_{0si}, \quad \pi_{1si} = \beta_{10i} + r_{1si}, \tag{5}$$

where β_{00i} represents the average outcome across series for participant i in the absence of treatment and β_{10i} is the average change in the outcome measures after the introduction of treatment to participant i. We assume that the level-2 error terms r_{0si} and r_{1si} are multivariate normally distributed with mean $(0, 0)'$ and covariance matrix $\mathbf{\Omega} = \begin{bmatrix} \omega_0^2 & \omega_{10} \\ \omega_{10} & \omega_1^2 \end{bmatrix}$. Finally, considering that the five participants have different levels of disability and communication skills, we assume that the average series-level outcome measures may vary across participants, and that the average series-level treatment effects vary across participants. These assumptions are expressed by the model

$$\beta_{00i} = \gamma_{000} + u_{00i}, \quad \beta_{10i} = \gamma_{100} + u_{10i}, \tag{6}$$

where γ_{000} is the overall average level of outcome measures in the absence of treatment and γ_{100} is the overall average treatment effect across series and across participants. We assume that the between-participant error terms u_{00i} and u_{10i} are multivariate normally distributed with mean $(0, 0)'$ and covariance matrix $\mathbf{T} = \begin{bmatrix} \tau_0^2 & \tau_{10} \\ \tau_{10} & \tau_1^2 \end{bmatrix}$.

We emphasize that the model given in Equations (4) through (6) is not the only possible specification for a replicated multiple baseline across behaviors or settings. In practice, the analyst will need to choose a model to capture the important features of the data. For instance, in other replicated multiple baseline design studies, it is possible that the data demonstrate time trends during the baseline or intervention phases. In such cases, the model could be expanded by adding a time variable and an interaction term between the treatment indicator and time at the first level of the analytic model. One would then need to consider whether the added first level coefficients vary across behaviors or across

participants. We demonstrate models that involve time trends in examples presented in subsequent sections. Another possibility is that, for some single-case studies, the assumption that treatment effects vary across behaviors or participants might not apply. Here, one could modify the model by assuming that the treatment effects are constant across behaviors or across participants. Generally, we recommend researchers using visual analysis to guide the initial model specification and then conduct preliminary analyses, including information criteria and likelihood ratio tests, to reach an appropriate model specification. We also recommend that researchers conduct and report sensitivity analyses to investigate the extent to which effect size estimates vary across different plausible model specifications.

Effect size definition

Given a model specification, a BC-SMD can be defined by considering a hypothetical between-group design in which the treatment is introduced after time A and outcomes are measured at time B. We assume that this hypothetical design would involve recruiting participants from the same population as those who took part in the single-case design, sampling a specific behavior or setting for each participant, and assessing the outcome for that behavior or setting. With this procedure, the numerator of the BC-SMD effect size represents the average effect of intervention, taken across behaviors and across participants, and the denominator of the BC-SMD includes the variation in the outcome across participants and across behaviors at a fixed point in time, in the absence of treatment. For the model given in Equations (4) through (6), the effect size parameter corresponds to

$$\delta_{AB} = \frac{\gamma_{100}}{\sqrt{\tau_0^2 + \omega_0^2 + \sigma^2}}. \tag{7}$$

The supplementary materials include a formal derivation of (7). Because the model in Equation (4) assumes stable baseline and treatment phases, the effect of the treatment is constant and the variation in the outcome is stable across time points. Consequently, δ_{AB}

does not depend on hypothetical time-points A or B for this particular model specification.

Effect size estimation

In order to calculate the BC-SMD in Equation (7), we need to set the P constants for fixed effect parameters and R constants for variance components in Equation (1). Based on the hierarchical model in Equation (4) to (6), the fixed effect parameters are the overall average outcome in the baseline phase γ_{000} and the overall average treatment effect γ_{100}. Thus, the constants for the fixed effect parameters are $p_1 = 0$ and $p_2 = 1$. The numerator of the BC-SMD is then calculated as $\sum_{a=1}^{2} p_a \gamma_a = \gamma_{100}$. The variance components are the between-participant intercept variance τ_0^2, the between-participant covariance between participant intercept and treatment effect τ_{10}, the between-participant treatment variance τ_1^2, the between-series intercept variance ω_0^2, the between-series covariance between the average outcome and treatment effect ω_{10}, the between-series treatment effect variance ω_1^2, the auto-correlation ϕ, and the within-series residual variance σ^2. The R constants for the variance components are thus $r_1 = r_4 = r_8 = 1$ and $r_2 = r_3 = r_5 = r_6 = r_7 = 0$ so that $\sqrt{\sum_{b=1}^{8} r_b \theta_b} = \sqrt{\tau_0^2 + \omega_0^2 + \sigma^2}$. Substituting the REML estimates for the fixed effect parameters and variance components and applying Equation (2) and (3) gives the small sample bias-corrected BC-SMD effect size estimate and standard error.

We used the *nlme* package in R (Pinheiro et al., 2019) to fit the hierarchical model and obtain the REML estimation for fixed effect parameters and variance component parameters. As shown in Table 1, the overall average treatment effect is $\hat{\gamma}_{100} = 11.052$ scale points and the total variation is $\hat{\tau}_0^2 + \hat{\omega}_0^2 + \hat{\sigma}^2 = 42.896$. We used the `g_mlm()` function in the *scdhlm* R package (Pustejovsky et al., 2020b) to calculate the BC-SMD effect size, the associated standard error, and the confidence intervals. The unadjusted BC-SMD is estimated as $\hat{\delta}_{AB} = \hat{\gamma}_{100} / \sqrt{\hat{\tau}_0^2 + \hat{\omega}_0^2 + \hat{\sigma}^2} = 1.687$ standard deviations. After applying the small sample corrected degrees of freedom, the adjusted BC-SMD is $g_{AB} = 1.659$ standard deviations, with a standard error of 0.306. A 95% symmetric CI for the adjusted effect size

Table 1
Model Estimates for Thiemann and Goldstein (2001) Data

Parameter	*Est*	*SE*
Variance Components		
Between-participant var ($\hat{\tau}^2$)	7.655	8.409
Participant-treatment cov ($\hat{\tau}_{10}$)	-6.523	7.344
Participant level treatment var ($\hat{\tau}_1^2$)	5.559	9.846
Between-series var ($\hat{\omega}_0^2$)	9.104	5.147
Series-treatment cov ($\hat{\omega}_{10}$)	0.058	5.055
Series level treatment var ($\hat{\omega}_1^2$)	14.492	9.820
Auto-correlation ($\hat{\phi}$)	0.078	0.059
Within-case var ($\hat{\sigma}^2$)	26.136	2.074
Total variance ($\hat{\tau}_0^2 + \hat{\omega}_0^2 + \hat{\sigma}^2$)	42.896	9.036
Fixed Effects		
Intercept ($\hat{\gamma}_{000}$)	6.062	1.523
Treatment ($\hat{\gamma}_{100}$)	11.052	1.614
Effect Size		
Unadjusted ($\hat{\delta}_{AB}$)	1.687	0.311
Adjusted (g_{AB})	1.659	0.306
Degrees of freedom (ν)	45.070	
Constant (κ)	0.246	
Log likelihood		
Log likelihood	-1129.597	

is [1.043, 2.276] using the symmetric t approach. The adjusted BC-SMD effect size estimate indicates that the average outcome in the treatment condition is 1.659 standard deviations above the average outcome in the baseline condition. This effect size can be interpreted as the average treatment effect that would be expected in a hypothetical between-group design where participants were first sampled, behaviors were then sampled for each participant, and outcomes were assessed at some (arbitrary) follow-up time.

Clustered Multiple Baseline Design

A second variation is the clustered multiple baseline across participants, in which each case is comprised of a group, such as students in a small group, classroom, or school, and where the intervention is implemented at the group level. In other words, the

intervention is introduced at different points in time for different groups while participants in the same group receive the intervention at the same time. Outcome data are collected at the individual level within each group. The structure of this design is analogous to a between-group, cluster-randomized experiment, in which intact groups of participants are assigned to receive an intervention or control condition. Just as in a cluster-randomized experiment, we seek to estimate a BC-SMD for the overall average effect of intervention across groups, accounting for the clustered nature of the intervention assignment process (Hedges, 2007; Taylor et al., 2021).

We illustrate the effect size estimation process using data from Bryant et al. (2018), who investigated the effects of an intensive mathematics intervention for second grade students with severe mathematics difficulties using a clustered multiple baseline design. The design involved 33 students from 12 groups in 5 schools. The intensive mathematics intervention was implemented for each group of students at the same time in one school but at different times for different schools. For each student, outcomes were measured with *Texas Early Mathematics Inventories-Aim Check* (TEMI-AC, Texas Education Agency/University of Texas System, 2009) during the baseline and intervention phases. In the primary study report, the authors aggregated student data to the group level for purposes of visual analysis and effect size calculations, which is a common practice with data from clustered multiple baseline designs. However, in order to examine the variability of the intervention effects across students and across groups, we use a model based on the raw, student-level data, without aggregation.[1] In this re-analysis, we exclude the maintenance phase data and focus on the difference in the outcomes between baseline phase and intervention phase. Student-level data are displayed in Figure 2.

[1] The primary study authors shared the student-level raw data with us for this re-analysis.

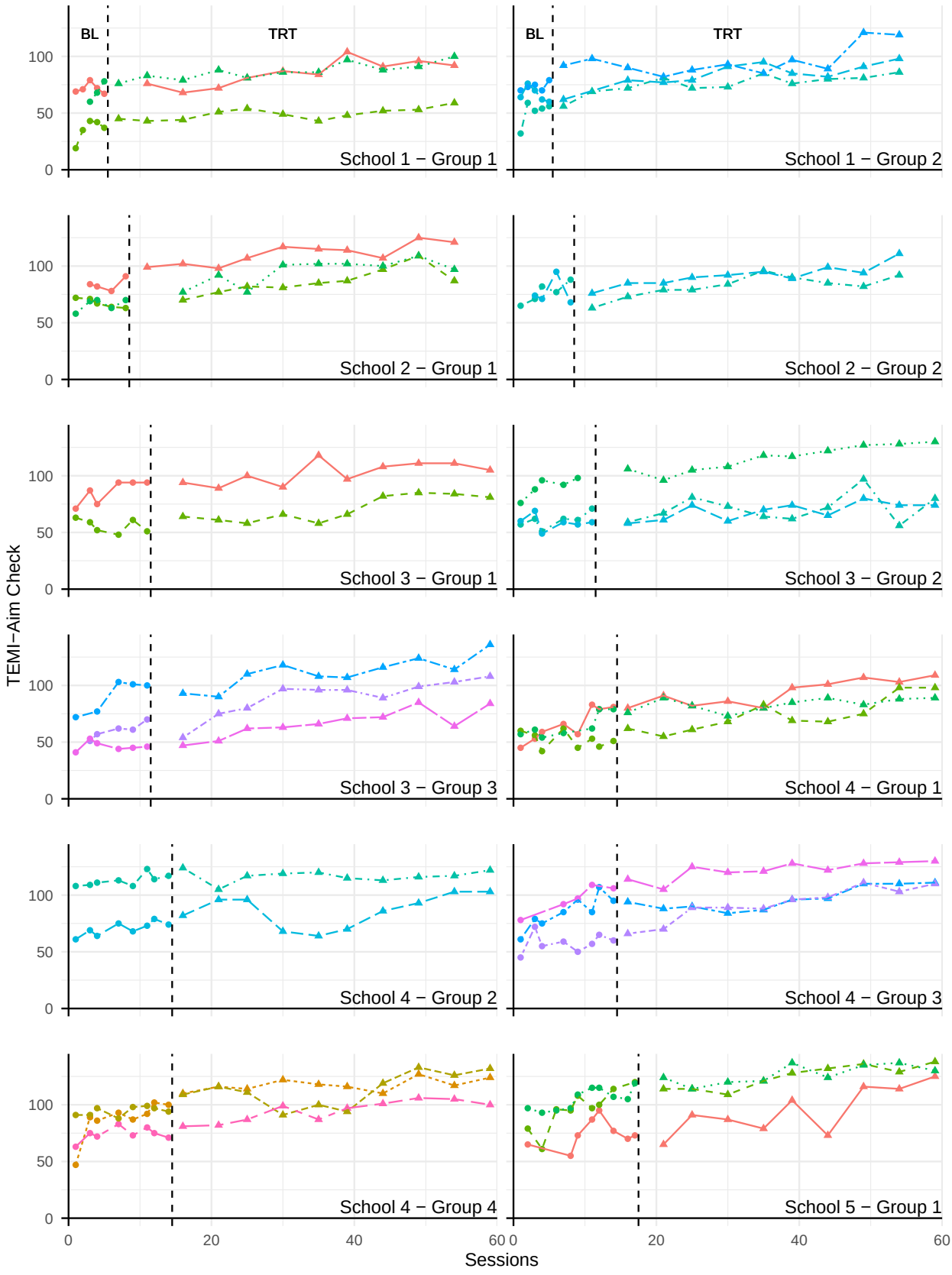

Figure 2

Clustered multiple baseline design data from Bryant et al. (2018).

Model specification

For estimating a BC-SMD effect size from a clustered multiple baseline design, we first need to specify a model that describes the observed data. In general, an appropriate hierarchical model for the clustered multiple baseline design will need to include levels for time-points nested within participants, for participants nested within groups, and for the groups to which treatment times are assigned. In order to estimate the overall average intervention effect across students and across groups for the Bryant et al. (2018) study, we use a three-level hierarchical model, with outcome measurements at the first level, students at the second level, and groups at the third level.[2]

Based on Figure 2, at the first level of the model, the baseline and treatment phase data seem to exhibit upward time trends and changes in the time trends between different treatment conditions for most students. At the second level of the model, the average outcome levels in the baseline phase, the baseline time trends, and the changes in time trends at the first level seem to vary across students (i.e., the second-level units) within the same group. Take the data from the first group in the first school as example. First, it is apparent that outcome levels differ for the three students. Second, one student's baseline data show no time trend while the other two students' baseline data indicate upward time trends but with different slopes. Third, it seems reasonable to conclude that the changes in the time trends from baseline to treatment phase vary across three students, with increase for one student and decrease for the other two students. At the third level of the model, it seems plausible that the average outcome levels vary from group to group. For instance, the average outcome values in the second group of school 4 or in the first group of school 5 are above 90 while the average outcome in the first group of school 1 is below 60. It is

[2] In this study, groups of students are actually nested within schools, which could be represented in the model as a further level (groups nested within schools). We ignore this aspect of the data for simplicity of illustration and because the limited number of groups within each school (ranging from one to four) makes it difficult to separate school-level variation from group-level variation. Thus, we use a three-level model specification.

difficult to decide whether other level-2 coefficients vary across groups based on visual analysis. We conducted preliminary analyses and the results supported the assumption that they are constant.

Let Y_{tij} denote the outcome measure at time t for student i in group j, where $t = 1, \ldots, N$, $i = 1, \ldots, m_j$, and $j = 1, \ldots, J$. Let T_j denote the time after which treatment is implemented for all students in group j. Let $I(t > T_j)$ be a dummy variable for treatment assignment, which equals 0 during the baseline phase and equals 1 during the treatment phase. Both the visual inspection described above and further preliminary analysis (see Section S2.1 of the supplementary materials) suggest that there might be time trends during the baseline phase and treatment phase for some students. We therefore use a three-level model that includes linear time trends in each phase and where time is centered at time C. The first-level model for the outcome measurements of student i in school j is then

$$Y_{tij} = \pi_{0ij} + \pi_{1ij} I(t > T_j) + \pi_{2ij}(t - C) + \pi_{3ij}\left[(t - T_{ij}) \times I(t > T_j)\right] + e_{tij}, \tag{8}$$

where π_{0ij} is the average level of outcome at time $t = C$ in the absence of treatment and π_{2ij} is the linear change in the outcome per session in the absence of treatment. The π_{1ij} represents the immediate change in the outcomes for student i in group j when treatment is introduced and π_{3ij} is the additional change in the outcome per session due to the implementation of treatment. Considering the repeated measurement of outcomes over time and conventions for statistical analysis of SCDs data, we also assume that the residual term e_{tij} has mean zero, variance σ^2, and first-order auto-correlation ϕ.

Second, based on the visual analysis, we assume that the student-specific parameters π_{0ij}, π_{2ij}, and π_{3ij} vary across students. Thus, the second-level model can be expressed as

$$\pi_{0ij} = \beta_{00j} + r_{0ij}, \quad \pi_{1ij} = \beta_{10j}, \quad \pi_{2ij} = \beta_{20j} + r_{2ij}, \quad \pi_{3ij} = \beta_{30j} + r_{3ij}, \tag{9}$$

where β_{00j} represents the average student-level outcome measures in group j in the absence of treatment and β_{20j} is the average change in the outcome per session across students in group j in the absence of treatment. The coefficient β_{10j} represents the average student-level immediate change in the outcome measures after the introduction of treatment in group j and β_{30j} is the additional average change in the outcome per session due to the treatment in group j. We assume that these error terms $(r_{0ij}, r_{2ij}, r_{3ij})'$ are multivariate normally distributed with mean $(0, 0, 0)'$ and covariance matrix

$$\mathbf{\Omega} = \begin{bmatrix} \omega_0^2 & \omega_{20} & \omega_{30} \\ \omega_{20} & \omega_2^2 & \omega_{32} \\ \omega_{30} & \omega_{32} & \omega_3^2 \end{bmatrix}.$$

Third, we assume that the average student-level outcome measures vary across groups but that the average time trends are equivalent across the groups. The third-level model is then given by

$$\beta_{00j} = \gamma_{000} + u_{00j}, \quad \beta_{10j} = \gamma_{100}, \quad \beta_{20j} = \gamma_{200}, \quad \beta_{30k} = \gamma_{300}, \tag{10}$$

where γ_{000} is the overall average level of outcome and γ_{200} is the overall average change in the outcome per measurement occasion, in the absence of treatment. The parameter γ_{100} is the overall average immediate change in the outcome after the introduction of treatment and γ_{300} is the overall average additional change in the outcome per session due to the treatment. The error term u_{00j} is the deviation from the overall average outcome from the average outcome in group j and we assume that $u_{00j} \sim N(0, \tau^2)$.

The model in Equations (8) to (10) is just one of many possible specifications that account for the data structure of a multi-level multiple baseline design. For data series that demonstrate different patterns in terms of the average outcome level, time trends, or variability, one might consider a different model specification. For example, if the data do

not exhibit baseline and intervention time trends at the individual level, one could remove the time variable and the interaction term between treatment and time in Equation (8). One might add or remove random slopes at the second or third level of the model based on the data pattern or knowledge of how individual participants were grouped. Considering the complexity of model specification, we expect that researchers will need to draw on theoretical knowledge of the intervention and outcome, visual analysis of the SCD data, and statistical strategies for model selection for determining an appropriate model.

Effect size definition

Based on our selected model specification for clustered multiple baseline design, a BC-SMD can be defined by considering a hypothetical cluster-randomized experiment where the treatment is assigned at the group level. We assume that the treatment is introduced after time A and outcomes are measured at time B. In such a hypothetical design, a standardized mean difference effect size would typically be defined as an overall average effect at follow-up time B, scaled by the total variation in the outcome at time B in the absence of treatment (Hedges, 2007). In our model specification, the numerator of the effect size corresponds to the initial average effect γ_{100} plus the additional change in the outcome due to treatment over the time from A to B, $\gamma_{300}(B-A)$. The denominator of the effect size includes the variation across groups, the variation across participants within groups, and idiosyncratic within-participant variation as of time B. Thus, the BC-SMD effect size corresponds to the combination of the model's parameters

$$\delta_{AB} = \frac{\gamma_{100} + \gamma_{300}(B-A)}{\sqrt{\tau_0^2 + \omega_0^2 + 2(B-C)\omega_{20} + (B-C)^2\omega_2^2 + \sigma^2}}. \tag{11}$$

Section S2 of the Supplementary Materials includes a formal derivation. Note that the denominator of Equation (11) depends on the follow-up time B because the model includes baseline time trends that vary across students (see Equation (9)). If time is centered at constant B by setting $C = B$, the effect size simplifies to

$\delta_{AB} = [\gamma_{100} + \gamma_{300}(B - A)] / \sqrt{\tau_0^2 + \omega_0^2 + \sigma^2}$.

Effect size estimation

Just as with the replicated multiple baseline design across behaviors, we need to determine the P constants for fixed effect parameters and R constants for variance components in Equation (1) in order to calculate a BC-SMD for the hierarchical model that we have specified. Based on the model specification in Equation (8) to (10), the constants for the fixed effect parameters are $p_1 = 0$, $p_2 = 1$, $p_3 = 0$, and $p_4 = B - A$, so that the numerator of the BC-SMD is $\sum_{a=1}^{4} p_a \gamma_a = \gamma_{100} + \gamma_{300}(B - A)$. The variance components include the group-level intercept variance τ^2; the student-level variances and covariances ω_0^2, ω_{20}, ω_2^2, ω_{30}, ω_{32}, ω_3^2; the auto-correlation ϕ, and the within-student residual variance σ^2. The R constants for the variance components are $r_1 = r_2 = 1$, $r_3 = \cdots = r_8 = 0$, and $r_9 = 1$, so that the denominator of the BC-SMD can be calculated as $\sqrt{\sum_{b=1}^{9} r_b \theta_b} = \sqrt{\tau^2 + \omega_0^2 + \sigma^2}$. Substituting the REML estimates for the fixed effect parameters and variance components and applying the small sample correction gives the bias-corrected BC-SMD effect size estimate and standard error.

We used the REML estimates of the fitted model to calculate BC-SMD effect size. We set the hypothetical treatment time to $A = 4$, which corresponds to the last observation of the shortest baseline phase of any of the groups. We set the hypothetical follow-up time to $B = 21$ so that the effects size estimate represents the effect of $B - A = 17$ days of mathematics intervention, which corresponds to the length of the shortest treatment phase of any of the groups. We centered the trends at day $C = 21$ to simplify the effect size calculation. As shown in Table 2, the numerator of the unadjusted effect size is the average treatment effect at measurement session 21, calculated as $\hat{\gamma}_{100} + \hat{\gamma}_{300}(21 - 4) = -0.222 - 0.794 \times 17 = -13.721$. The denominator of the unadjusted effect size is a linear combination of the variance components at session 21, calculated as $(\hat{\tau}^2 + \hat{\omega}_0^2 + \hat{\sigma}^2) = 3.975 + 559.647 + 59.075 = 622.698$. Thus, the unadjusted BC-SMD is

Table 2
Model Estimates for Bryant et al. (2018) Data

Parameter	*Est*	*SE*
Variance Components		
Between-group var($\hat{\tau}^2$)	3.975	43.148
Between-student var ($\hat{\omega}_0^2$)	559.647	181.119
Student-trend cov ($\hat{\omega}_{20}$)	16.498	7.175
Student level trend var ($\hat{\omega}_2^2$)	0.778	0.352
Student-Trt. × Trend cov ($\hat{\omega}_{30}$)	-18.049	7.850
Student level Trend-Trt. × Trend cov ($\hat{\omega}_{32}$)	-0.814	0.380
Student level Trt. × Trend var ($\hat{\omega}_3^2$)	0.886	0.417
Auto-correlation ($\hat{\phi}$)	0.328	0.103
Within-student var ($\hat{\sigma}^2$)	59.075	4.256
Total variance ($\hat{\tau}^2 + \hat{\omega}_0^2 + \hat{\sigma}^2$)	622.698	177.012
Fixed Effects		
Intercept ($\hat{\gamma}_{000}$)	93.146	5.038
Treatment ($\hat{\gamma}_{100}$)	-0.222	1.311
Trend ($\hat{\gamma}_{200}$)	1.346	0.235
Trt. × Trend ($\hat{\gamma}_{300}$)	-0.794	0.244
Trt. effect after 21 days $\left(\sum_{a=1}^{P} p_a \hat{\gamma}_a\right)$	-13.721	4.743
Effect Size		
Unadjusted ($\hat{\delta}_{AB}$)	-0.550	0.215
Adjusted (g_{AB})	-0.533	0.208
Degrees of freedom (ν)	24.750	
Constant (κ)	0.190	
Log likelihood		
Log likelihood	-1947.094	

$\hat{\delta}_{AB} = (\hat{\gamma}_{100} + 17 \times \hat{\gamma}_{300})/\sqrt{\hat{\tau}^2 + \hat{\omega}_0^2 + \hat{\sigma}^2} = -.550$ with a standard error of .215. The small-sample adjusted BC-SMD is very similar, $g_{AB} = -.533$, with a standard error of .208 and a 95% CI based on symmetric t approximation of [-0.962, -0.104]. The adjusted BC-SMD effect size estimate means that the average outcome of participants across groups in the treatment condition at time B is .533 standard deviations below the average outcome in the baseline condition. It represents the average effect at time 21 of introducing treatment after time 4 in a hypothetical between-group cluster-randomized experiment

where each group of participants was assigned to receive an intervention or to a control condition.

Although visual inspection of the data suggests a model with varying intercept across students and groups and varying time trends and treatment-by-trend interaction across students, we conducted sensitivity analyses using several other model specifications. One alternative model was based on assuming stable baselines without time trends. Another assumed that baseline time trends and treatment-by-time interactions were constant across participants and across groups, rather than allowing them to vary as in Equation (9). Full sensitivity analyses results are presented in Section S2.1 of the Supplementary Materials. We used the AIC, BIC, and likelihood ratio tests for model comparison. All of these indicators suggested that the model specified in Equation (8) to (10) fit the data best, which is consistent with the visual inspection. The BC-SMD effect size estimates based on the models assuming no baseline time trends were positive, with bias-corrected estimates of about 0.60. In contrast, estimates based on models allowing for linear time trends in the baseline and intervention phases were all negative, with values ranging from -0.766 to -0.533. The sensitivity of the BC-SMD estimates from this study highlights the degree to which effect size estimates from SCDs are contingent on modeling assumptions, as well as the importance of using visual inspection and domain knowledge to inform model specification.

As a further sensitivity analysis, we also estimated effect sizes after aggregating the student-level data to the group level (as done in the primary study report). We then calculated the BC-SMD effect size based on a hierarchical model for the aggregated measurements nested within 12 groups, using the simpler two-level approach described by Pustejovsky et al. (2014). Supplementary Materials Section S2.2 include complete results from this analysis. The small sample bias-corrected BC-SMD effect size estimate based on a model with similar specification as the three-level model was -0.951 (SE = 0.435, 95% asymmetric CI: [-1.834, -0.215]). The estimate based on the aggregated data has the same

negative sign as those based on the student-level data, but was substantially larger in absolute magnitude. Analyzing the aggregated data reduces the scale parameter in the denominator of the BC-SMD and thus overstates the intervention effect.

Overall, our effect size estimates indicate that the intensive mathematics intervention, as implemented in this study, has a detrimental average effect on student math performance based on models that account for linear time trends, but a beneficial average effect based on models that assume baseline stability. In the original paper, the authors calculated NAP effect sizes, which indicated no overlap for most of the groups. The NAP estimates suggest a positive intervention effect, consistent with the BC-SMD effect size estimates without assuming time trends. However, based on the visual analysis and model comparisons, we put greater stock in the effect size estimates based on models that include time trends.

Multivariate Multiple Baseline Design Across Participants

A third variation of the multiple baseline design is the multivariate multiple baseline across participants, where a multiple baseline across participants is applied to evaluate the intervention effects for not just one, but several, distinct outcome measures. Like other multiple baseline designs, this design involves staggering the introduction of intervention, so that different participants receive intervention beginning at different points in time. The unique feature of the design is that it involves collecting *multiple* outcome measurements at each time-point for each participant. As a result, the outcomes may be correlated, leading to dependence in the effect size estimates.

With currently available methods for estimating BC-SMDs, data from a multivariate multiple baseline design could be analyzed by fitting separate models for each of the outcome measures. This approach amounts to treating the study as several distinct multiple baseline designs across participants, even though the designs all involve the same participants and interventions and all take place concurrently. We instead propose a

different approach, which entails analyzing all of the data within a single model that allows for dependence across outcome measurements. We can then estimate multiple BC-SMD effect sizes from the model (i.e., one effect size per outcome) and assess not just the standard errors of each estimate, but also the covariance among the effect size estimates.

Using a single, multivariate model has several potential advantages, which are similar to the advantages of multivariate methods in other contexts (Snijders & Bosker, 2011, Chapter 16, pp. 282-288). First, applying one single model provides a way to account for the correlations between pairs of the distinct outcome measures, which is useful if the effect size estimates are all going to be analyzed in a meta-analysis. Second, because the correlations among outcome measures are taken into account, the approach allows us to make comparisons of effect sizes across outcomes, such as estimating a difference between two effect sizes. Third, using a single model allows the model to borrow information across outcomes, potentially leading to more precise estimates than those generated from using separate models for each outcome.

Calder et al. (2020) is an example of a multivariate multiple baseline across participants. The authors examined the effects of an explicit grammar instruction intervention on expressive morphosyntax (i.e., conjugation of past tense verbs) and grammaticality judgement for nine children with developmental language disorder. The nine participating students were randomly assigned to three distinct intervention starting points. Therefore, the intervention was introduced at different times across participants but at the same time for each of the outcomes measured on a given participant. The authors assessed expressive morphosyntax and grammaticality judgements for both trained and untrained verbs.[3] We analyze the data on untrained verbs to provide effect size measures describing more generalized effects of intervention. We used a bivariate model, where two outcome measures were percentage correct on expressive morphosyntax (i.e.,

[3] The researchers also assessed the expression and grammaticality judgement of third-person singular verbs and possessive *'s* as an extension and a control for the past tense verbs, respectively. For purposes of illustration, we limited the analysis to outcomes for the target behavior of past tense marking.

past tense production on untrained verbs) and grammaticality judgements about grammatical and ungrammatical sentences with untrained past tense verbs. The raw data are displayed in Figure 3, with columns corresponding to two distinct outcome measures and rows corresponding to different participants. For purposes of analysis, we organized the raw data in a long format where each row includes the outcome-specific response for one measurement occasion of one participant (Pustejovsky et al., 2020b).

The multivariate multiple baseline across participants reported by Calder et al. (2020) served as a pilot study for a larger, between-group randomized control trial of the same intervention (Calder et al., 2021). Outcomes in the between-group design included expressive morphosyntax and grammaticality judgement measures, just as in the multivariate multiple baseline design. Outcomes were assessed at the conclusion of a ten week intervention. The existence of such a between-group design provides further motivation for estimating BC-SMD effect sizes from Calder et al. (2020).

Model Specification

Generally speaking, to estimate a BC-SMD effect size based on a single hierarchical model for the data from a multivariate multiple baseline across participants, the model will describe the repeated measurements (nested within participants) at the first level and participants at the second level and will allow for the dependence among distinct outcomes. For the within-participant level (first-level) of the model, we first need to decide whether the baseline phase and treatment phase data indicate systematic time trends. Figure 3 indicates that baseline and treatment time trends appear to be present for most participants. For example, for the expressive morphosyntax outcome, *P1*, *P3*, *P9* seem to indicate a downward trend while *P4* and *P8* appear to have an upward trend in baseline, and most participants appear to have an upward trend in the treatment phase. For the grammaticaltiy judgement outcome, *P8* and *P6* seem to demonstrate a downward trend in the baseline phase and intervention phase, respectively.

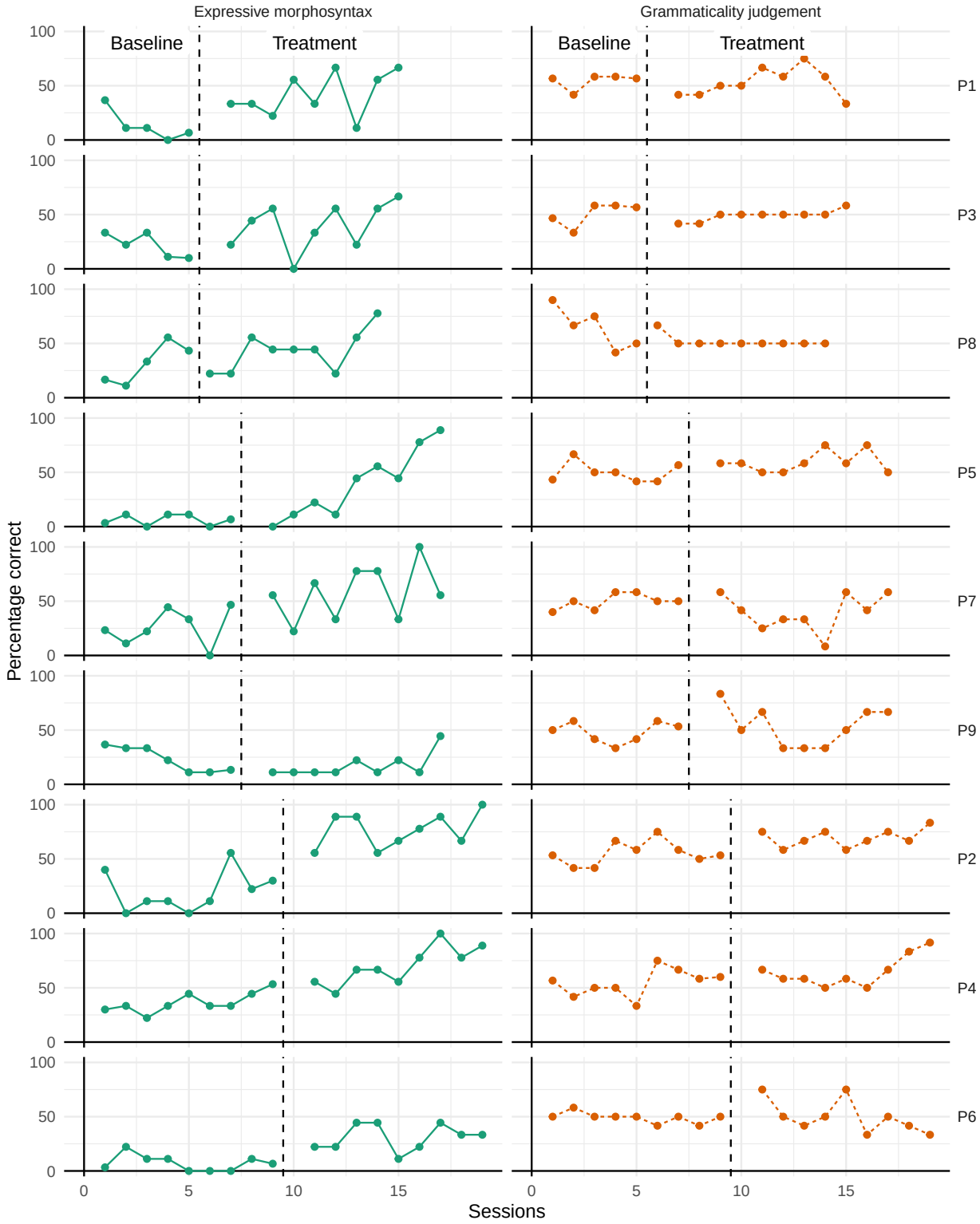

Figure 3

Multivariate multiple baseline design across participant data from Calder et al. (2020).

At the between-participant level (second-level) of the model, we need to make assumptions about whether the within-participant level coefficients vary across participants. Visual inspection suggests that the intercept and time trends in the baseline or intervention phase vary across participants. Second, we need to determine whether the baseline level of the outcome, treatment effects, and time trends are different for different outcome measures. The data in Figure 3 demonstrates different baseline levels, changes in percent correct, or baseline and intervention time trends across two outcome measures. Third, we need to consider the correlation structure among the two outcome measures at each point of time, as well as the heterogeneity for different outcome measures. The two outcome measures (expressive morphosyntax and grammaticality judgement) might be correlated because they were measured based on a common sample of participants and they both tested the participant's underlying understanding of past tense marking. However, the within-participant error variance might vary across outcomes because they were measured with different procedures. Thus, we use a model that allowed for correlation between two outcome measures and heterogeneous within-participant error variances across outcomes.

Based on the visual analysis above and further preliminary analysis (see Section S3.1 of the supplementary materials), we fit a two-level model that allows for different baseline levels, treatment effects, and time trends in the baseline and intervention phase across outcome measures for each participant, and for varying intercept and baseline time trends across participants. Let Y_{ti}^1 and Y_{ti}^2 indicate the percentage correct on the expressive morphosyntax and grammticality judgement of untrained verbs, respectively, for participant i at time t, where $t = 1, \ldots, N$ and $i = 1, \ldots, m$. The model for the k-th outcome for participant i at time t is

$$Y_{ti}^k = \beta_{0i}^k + \beta_{1i}^k I(t > T_i^k) + \beta_{2i}^k (t - C) + \beta_{3i}^k \left[(t - T_i^k) \times I(t > T_i^k) \right] + e_{ti}^k, \tag{12}$$

where β_{0i}^k represents the average level of the k-th outcome for participant i at time $t = C$ in

the absence of treatment. The β_{1i}^k represents the immediate change in the k-th outcome for participant i due to the implementation of treatment. The coefficient β_{2i}^k is the linear change in k-th outcome per measurement occasion for participant i in the absence of treatment and β_{3i}^k is the additional change in the k-th outcome per measurement occasion for participant i after introducing the treatment. Finally, $(e_{ti}^1, e_{ti}^2)'$ is the vector of within-participant residuals of the two outcomes, where we assume that the errors have means of zero and variance-covariance matrix $\Sigma = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$, where σ_1^2 and σ_2^2 represent the unique within-participant error variances for the expressive morphosyntax and grammaticality judgement errors, respectively, and ρ_{12} is the correlation between the within-participant errors.

At the participant level, we specify the model as

$$\beta_{0i}^k = \gamma_{00}^k + r_{0i}^k, \quad \beta_{1i}^k = \gamma_{10}^k, \quad \beta_{2i}^k = \gamma_{20}^k + r_{2i}^k, \quad \beta_{3i}^k = \gamma_{30}^k, \tag{13}$$

where γ_{00}^k represents the average level of the k-th outcome across all participants in the absence of treatment; γ_{10}^k is the average immediate change in the k-th outcome due to treatment, which is assumed to be constant across participants; γ_{20}^k is the average change in the k-th outcome per measurement occasion; and γ_{30}^k, is the time-by-treatment interaction, or average additional change in the k-th outcome as a result of treatment, assumed to be constant across participants. Finally, we assume that $(r_{0i}^k, r_{2i}^k)'$ follows a multivariate normal distribution with mean $(0^1 \dots 0^k, 0^1 \dots 0^k)'$ and covariance matrix $\mathbf{T}$. To simply the notation, we present the covariance matrix $\mathbf{T}$ for this example as follows:

$$\mathbf{T} = \begin{bmatrix} \mathbf{T}_{00} & \mathbf{T}_{02} \\ \mathbf{T}_{02}^T & \mathbf{T}_{22} \end{bmatrix},$$

where

$$\mathbf{T}_{00} = \begin{bmatrix} \tau_{00}^{1} & \tau_{00}^{1,2} \\ \tau_{00}^{1,2} & \tau_{00}^{2} \end{bmatrix}, \quad \mathbf{T}_{02} = \begin{bmatrix} \tau_{02}^{1,1} & \tau_{02}^{1,2} \\ \tau_{02}^{2,1} & \tau_{02}^{2,2} \end{bmatrix}, \quad \mathbf{T}_{22} = \begin{bmatrix} \tau_{22}^{1} & \tau_{22}^{1,2} \\ \tau_{22}^{1,2} & \tau_{22}^{2} \end{bmatrix},$$

and where τ_{pq}^{k} denotes the variance or covariance between the errors r_{pi}^{k} and r_{qi}^{k} that pertain to outcome $k = 1, 2$ and where $\tau_{pq}^{k,l}$ denotes the covariance between the errors r_{pi}^{k} and r_{qi}^{l} for different outcomes, so $k = 1, 2$ and $l = 3 - k$.

The model specification in Equation (12) and (13) is based on visual inspection of the raw data from this particular example of a multivariate multiple baseline across participants. One could change some of the model aspects to account for the properties of a different study that used the same design. For example, we may allow $\beta 1 i^{k}$ or β_{3i}^{k} to vary across participants. In other applications, we might assume a homogeneous within-participant variance so that $\text{Var}(e_{ti}^{k}) = \sigma^2$ and $\text{Cov}(e_{ti}^{k}, e_{ti}^{l}) = \sigma^2 \rho_{kl}$, or use a different correlation structure among outcome measures. In addition to the model specification strategies discussed in the previous applications, we recommend developing models for the multivariate multiple baseline design across participants by selecting a specification that fits each of the outcomes individually, then considering how to combine the specifications into a common model.

Effect size definition

For estimating BC-SMD effect sizes for distinct outcome measures in the multivariate multiple baseline across participants, we consider a hypothetical between-group design where participants were randomly assigned to control and intervention groups, intervention were introduced after time A, and responses for each outcome measure were collected at time B. The BC-SMD effect size for the k-th outcome measure is defined as the mean difference in the k-th outcomes at time B if the treatment were introduced after time A versus if the treatment were never introduced, scaled by the square root of the variation in the outcomes for k-th measure at time B if the treatment

were never introduced. The BC-SMD effect size for the k-th outcome measure based on the model specifications in Equation (12) and Equation (13) is therefore

$$\delta_{AB}^{k} = \frac{\gamma_{10}^{k} + \gamma_{30}^{k}(B - A)}{\sqrt{\tau_{00}^{k} + 2(B - C)\tau_{20}^{k} + (B - C)^2\tau_{22}^{k} + \sigma_{k}^{2}}}.$$

(see Supplementary Materials Section S3 for a formal derivation). When time is centered on constant B by setting $C = B$, the BC-SMD effect size for outcome k simplifies to

$$\delta_{AB}^{k} = \frac{\gamma_{10}^{k} + \gamma_{30}^{k}(\mathrm{B} - A)}{\sqrt{\tau_{00}^{k} + \sigma_{k}^{2}}}. \tag{14}$$

Because the effect sizes are defined based on a single model encompassing multiple outcomes, we will be able to borrow information across outcomes in order to estimate each effect size, as well as to estimate the sampling covariance among the effect sizes.

Effect size estimation

To calculate the BC-SMD effect sizes defined in Equation (14) using the parameters of the hierarchical linear model, the P constants for fixed effect parameters and R constants for variance components need to be determined for each outcome. According to the model specification in Equation (12) and (12), the fixed effect parameters are the outcome specific average in the absence of treatment ($\gamma_{00}^{1},\ \gamma_{00}^{2}$), the average immediate change for the k-th outcome ($\gamma_{10}^{1},\ \gamma_{10}^{2}$), the average change in the k-th outcome per measurement occasion ($\gamma_{20}^{1},\ \gamma_{20}^{2}$), and the additional change for the k-th outcome ($\gamma_{30}^{1},\ \gamma_{30}^{2}$). Thus, for the expressive morphosyntax outcome, the P constants are $p_{3}^{1} = 1$ and $p_{7}^{1} = B - A$ with other constants equal to 0. For the grammaticality judgement outcome, the P constants are $p_{4}^{2} = 1$ and $p_{8}^{2} = B - A$ with others equal to 0. The numerator of the BC-SMD is thus calculated as $\sum_{a=1}^{8} p_{a}^{1}\gamma_{a}^{1} = \gamma_{10}^{1} + \gamma_{30}^{1}(B - A)$ for the expressive morphosyntax, and $\sum_{a=1}^{8} p_{a}^{2}\gamma_{a}^{2} = \gamma_{10}^{2} + \gamma_{30}^{2}(B - A)$ for the grammaticality judgement.

The variance components are the between-student intercept variance for expressive

morphosyntax (τ^1_{00}), the between-student covariance of outcome specific intercepts ($\tau^{1,2}_{00}$), the between-student intercept variance for grammaticality judgement (τ^2_{00}), the between-student covariance of expressive morphosyntax trend and intercept ($\tau^{1,1}_{02}$), the between-student covariance of expressive morphosyntax trend and grammticality judgement intercept ($\tau^{2,1}_{02}$), the between-student covariance of grammaticality judgement trend and expressive morphosyntax intercept ($\tau^{1,2}_{02}$), the between-student covariance of grammaticality judgement trend and intercept ($\tau^{2,2}_{02}$), the outcome specific between-student trend variance ($\tau^1_{22},\ \tau^2_{22}$), the between-student trend covariance between two outcomes ($\tau^{1,2}_{22}$), the within-student correlation between two outcomes (ρ_{12}), and the within-student variance for expressive morphosyntax (σ^2_1) and for grammaticality judgement (σ^2_2). Thus, for the expressive morphosyntax outcome, the R constants are $r^1_1 = r^1_{12} = 1$ with other constants equal to 0. For the grammaticality judgement outcome, the R constants are $r^2_3 = r^2_{13} = 1$ with others equal to 0. The denominator of the BC-SMD effect size is thus $\sqrt{\sum_{b=1}^{13} r^1_b \theta^1_b} = \sqrt{\tau^1_{00} + \sigma^2_1}$ for expressive morphosyntax and $\sqrt{\sum_{b=1}^{13} r^2_b \theta^2_b} = \sqrt{\tau^2_{00} + \sigma^2_2}$. We estimated the BC-SMD effect sizes using the REML estimates from the fitted hierarchical model and applied the small sample correction for obtaining bias-corrected BC-SMD effect sizes and standard errors for each outcome measure.

Table 3 reports the model estimates and standard errors for the expressive morphosyntax and grammaticality judgement outcome (in columns). Following the default in the *scdhlm* web application, we set $A = 5$ because it was the last measurement occasion of the shortest baseline phase. We set $B = 14$ because it corresponded to the shortest treatment duration across participants and because it was similar to the ten-session intervention examined in the subsequent between-group design by Calder et al. (2021). The effect size estimates thus represent the effect of $B - A = 9$ sessions of the intervention. We centered time at $C = 14$ to simplify the effect size calculation. The numerator of the unadjusted effect size is calculated with $\hat{\gamma}^k_{10} + \hat{\gamma}^k_{30}(14 - 5)$ for the k-th outcome measure. The denominator of the unadjusted effect size is calculated with $(\hat{\tau}^k_{00} + \hat{\sigma}^2_k)$. Thus, the

Table 3
Model Estimates for Calder et al. (2020) Data.

Parameter	Expressive morphosyntax *Est. (SE)*	Grammaticality judgement *Est. (SE)*
Variance Components		
Between-student var ($\hat{\tau}^1_{00}, \hat{\tau}^2_{00}$)	314.007 (174.175)	66.218 (42.621)
Student level cov btw exp and ($\hat{\tau}^{1,2}_{00}$)		78.923 (66.609)
Cov btw exp-trend and ($\hat{\tau}^{1,1}_{02}, \hat{\tau}^{2,1}_{02}$)	23.891 (14.809)	7.825 (6.176)
Cov btw grm-trend and ($\hat{\tau}^{1,2}_{02}, \hat{\tau}^{2,2}_{02}$)	6.455 (7.093)	6.523 (4.457)
Trend var ($\hat{\tau}^1_{22}, \hat{\tau}^2_{22}$)	2.294 (1.457)	0.697 (0.523)
Trend cov btw exp and ($\hat{\tau}^{1,2}_{22}$)		0.743 (0.663)
Corr between exp and ($\hat{\rho}_{12}$)		-0.14 (0.088)
Within-student var ($\hat{\sigma}^2$)	234.333 (29.75)	134.412 (17.056)
Total variance ($\hat{\tau}_{00} + \hat{\sigma}^2$)	548.34 (175.989)	200.629 (45.05)
Fixed Effects		
Intercept ($\hat{\gamma}_{00}$)	22.041 (10.852)	55.038 (6.772)
Treatment ($\hat{\gamma}_{10}$)	0.999 (5.296)	-1.512 (3.869)
Trend ($\hat{\gamma}_{20}$)	0.203 (1.02)	0.274 (0.668)
Trt. $\times$ Trend ($\hat{\gamma}_{30}$)	3.827 (1.106)	0.108 (0.79)
Trt. effect after 14 sessions $\left(\sum_{a=1}^{P} p_a \hat{\gamma}_a\right)$	35.442 (11.017)	-0.539 (7.561)
Effect Size		
Unadjusted ($\hat{\delta}$)	1.514 (0.56)	-0.038 (0.548)
Adjusted (g)	1.454 (0.538)	-0.037 (0.537)
Degrees of freedom (ν)	19.416	39.668
Constant (κ)	0.47	0.534
Log likelihood		
Log likelihood	-1161.705	-1161.705

unadjusted effect size is estimated as $\hat{\delta}^k_{AB} = (\hat{\gamma}^k_{10} + 9\hat{\gamma}^k_{30})/(\hat{\tau}^k_{00} + \hat{\sigma}^2_k)$ for the k-th outcome and $\hat{\delta}^1_{AB} = 1.514$ for expressive morphosyntax and $\hat{\delta}^2_{AB} = -.038$ for grammaticality judgement. The small-sample adjusted effect size is $g^1_{AB} = 1.454$ with a standard error of .538 for expressive morphosyntax and $g^2_{AB} = -.037$ with a standard error of .537 for grammaticality judgement. A symmetric 95% CI is [0.329, 2.579] for expressive morphosyntax and [-1.124, 1.049] for grammaticality judgement. The adjusted BC-SMD

effect size estimate for the expressive morphosyntax indicates that the average outcome in the treatment condition at time B is 1.454 standard deviations above the average outcome in the baseline condition. It describes the treatment effect at time $B = 14$ of introducing treatment at time $A = 5$ in a hypothetical between-group design. The adjusted effect size estimate for grammaticality judgement can be interpreted the same way.

We also estimated BC-SMD effect sizes by analyzing the data for each outcome separately. The Supplementary Section S3.3 reports results for each outcome. When using separate models, the biased corrected BC-SMD effect size estimates are $g^1_{AB} = 1.423$ (SE = .544) for expressive morphosyntax and $g^2_{AB} = -.113$(SE = .636) for grammaticality judgement. For the expressive morphosyntax, the standard error of the bias-corrected BC-SMD from the separate model is similar to (slightly greater than) the corresponding standard error from the multivariate model. However, for the gramaticality judgement, the standard error of the bias-corrected BC-SMD in the separate model is about 18% greater than that of the single model. This illustrates that the multivariate model can provide more precise effect size estimates than the separate models.

In Supplementary Materials Section S4, we provide mathematical details about how to estimate the sampling covariances among BC-SMD effect size estimates, which are needed for assessing the uncertainty of *differences* between effect sizes and when including the effect size estimates in a meta-analysis. The sampling covariance between two BC-SMD effect size estimates in this example is -0.01, implying a correlation of -0.05. The negative correlation between the BC-SMD effect size estimates is due to the negative correlation between within-participant errors of the two outcomes. Although the BC-SMD effect size estimates are similar in the single multivariate model and the separate models and the correlation between two BC-SMD estimates is small, we still prefer the results based on the single multivariate model, given the general advantages discussed in the previous context.

Discussion

Pustejovsky et al. (2014) proposed a general framework for defining and estimating BC-SMD effect sizes for across-participant multiple baseline designs. In practice, some SCDs have a more complex data structure, to which the methods described in Pustejovsky et al. (2014) cannot be readily applied. In the current study, we discuss the model specification, BC-SMD effect size definition, and estimation methods for three extensions of the multiple baseline design, and we demonstrate their application with data from published SCD studies. First, in designs that involve replicating a multiple baseline across behaviors or settings for several participants, BC-SMD methods can be used for estimating an average effect size across behaviors and across participants. Second, in the clustered multiple baseline design where participants are nested within groups and full groups are assigned to staggered intervention starting times, BC-SMD methods can be used to estimate an overall average BC-SMD effect size that is in the metric of the SMD effect size from a hypothetical cluster-randomized trial conducted with the same groups of participants (cf. Hedges, 2007). In contrast to applying previously proposed methods to group-level aggregated data, our approach models individual-level data in order to account for within-group and between-group variation in the outcomes. Third, in the multivariate multiple baseline design across participants, we describe an approach that models dependence across several distinct outcome measures, so that multiple effect sizes can be defined and estimated based on a common model. We can also estimate the sampling covariances of the effect size estimates, which is useful for meta-analysis of multivariate effect sizes.

We have focused on extending the BC-SMD framework to three further variations of the multiple baseline design. However, the BC-SMD framework is not limited to these specific designs, and the same principles can be applied to other types of designs that researchers might use in practice. For example, researchers might use a multiple baseline across behaviors and replicate the design across participants nested in classrooms or

schools. In such a design, a BC-SMD effect size could be estimated by fitting a four-level hierarchical linear model. As another example, researchers might use a clustered multiple baseline design but collect several related outcome measures. In such a design, our modeling approach could be further extended to handle both the multi-level structure and the multivariate outcomes.

Model specification

In extending the BC-SMD framework to these more complex cases, our primary focus has been on how to define and estimate effect sizes, given an appropriate model specification. A major outstanding challenge for applying this framework (and for modeling of SCD data more generally) is how to choose a model specification for a given application (Li et al., 2021). In the hierarchical models we have applied, one needs to determine which of the lower-level parameters of the model vary across higher level units. Different model assumptions can result in different fixed effects or variance components estimates, leading to different BC-SMD effect size estimates. In the applications we have reported here, we have used visual inspection to guide the main aspects of model specification. We also recommend starting from a simple and constrained model, then relaxing some of the constraints, and conducting model comparison. Researchers might use information criteria for model comparison. However, the asymptotic approximations behind information criteria might be violated in SCDs due to the small number of top-level units (Gurka, 2006), or different comparison criteria might not be consistent. Thus, in using such criteria, it is important to ensure that statistical analyses are consistent with visual inspection and knowledge of the population and intervention under study. As in any quantitative analysis, it is critical for researchers using BC-SMD effect sizes to be explicit about the model specification and selection process and, when possible, to provide theoretical rationales for their choices.

Defining effect sizes

Another challenge for applying the BC-SMD framework is deciding which variance components should be used for scaling the effect size and which treatment times and follow-up times should be used for defining the effect size. Because the treatment effects are standardized with a scale parameter that includes the residual variance at the first level and the variance components at the second or higher levels of the hierarchical model, further judgment is needed for choosing an appropriate combination of variance components. In the context of conducting a meta-analysis of several multiple baseline designs, we recommend considering the features of all included studies and, to the extent possible, making modeling choices that are consistent across studies. For instance, researchers could examine the range of treatment phase lengths across all included studies and choose treatment and follow-up time-points so that the hypothetical treatment duration corresponds to a typical treatment duration (e.g., near the middle of the range). Researchers can also augment their calculations with sensitivity analysis based on longer or shorter treatment durations, where the range of durations examined is informed by the range of follow-up times observed across studies. Of course, all of this requires that the timing of measurement occasions is defined in a clear and consistent fashion across the included studies. Improvements in the reporting of measurement occasion timing and related aspects of primary studies (Ledford et al., 2022) would facilitate better integration in research synthesis.

As another example, consider a clustered multiple baseline design where students are nested within schools, and where the BC-SMD could be defined by including the school-level variance components. If treatment is at the school level across most included studies, then we can include the school-level variance when calculating BC-SMD effect sizes. However, if most of the other included studies involve implementing an individual-level treatment in a single school, then the school-level variance cannot be estimated in those studies. In order to maintain comparability with these other studies, we would recommend

calculating a BC-SMD effect size for the clustered multiple baseline design that does *not* include school-level variance. Regardless of which variance components are used in the effect size, improvements in open data practices (Cook et al., 2021; Ledford et al., 2022) would facilitate effect size calculations for clustered multiple baseline designs.

In the clustered multiple baseline design example, we found that the BC-SMD effect size estimate based on the two-level model for the aggregated data was much larger than the estimate based on the three-level model for the raw data without aggregation. This is because the two-level model based on aggregated data had smaller variation that was used in the denominator of the effect size parameter, resulting in over-estimated BC-SMD effect size estimates. This over-estimation can be substantial–especially when group sizes are large. Thus, researchers would ideally use the raw data for model specification and effect size estimation. However, it might be difficult to access the raw data. In that case, it is theoretically possible to apply a correction to the effect size estimates of the aggregated data to receive the effects that would have been obtained if the analyses were based on the individual data. In between-group designs, this correction is a function of group size and intra-class correlation (see Chapter 3, Snijders & Bosker, 2011, pp. 17–26). In fact, the correction factor is larger for larger group sizes and larger intra-class correlation. The exact form of the correction factor for the clustered multiple baseline design is provided in the Section S2.3 of the supplementary materials. Although this correction is theoretically feasible and can be used for adjusting effect size estimates based on aggregated data, it requires empirical evidence about the intra-class correlation of the outcome, which might not be available in practice. Therefore, we strongly recommend researchers share raw data for clustered multiple baseline designs.

With the extension of BC-SMD methods to these variations of multiple baseline design, it is worth considering how to conduct meta-analysis of SCDs that involve different types of multiple baselines, especially when both individual-level and clustered multiple baseline designs are included in a meta-analysis. Researchers could include the BC-SMD

effect sizes from both variations 1) if the individual level data is available and BC-SMD effect sizes are estimated based on a three-level model that accounts for clustering in clustered multiple baseline design, or 2) if the individual level data is not available but the aggregation correction can be estimated and applied to effect size estimates in clustered multiple baseline design. Additionally, we recommend that researchers conducting meta-analyses of single-case designs use moderator analysis to further investigate differences in effect size for different types of SCDs.

Limitations

The BC-SMD methods that we have described are limited in several respects that are important to consider. First, the BC-SMD effect size index describes an *average* intervention effect across cases (e.g., participants) but does not reflect the change in the outcome within each individual participant (Odom et al., 2018). Because of its theoretical comparability to SMD effect sizes for group designs (which also describe an average effect), the BC-SMD effect size can be used in the research synthesis that includes both single-case and group design studies. However, if the aim of a synthesis is to examine variation in the outcome within each case or individual participant, to examine individual-level predictors of effects, or to summarize results for a review comprised exclusively of SCDs, researchers might it advantageous to use other methods, such as within-case effect sizes that describe individual-level intervention effects (Moeyaert et al., 2018; Pustejovsky & Ferron, 2017) or raw data synthesis methods (Moeyaert et al., 2017; Van den Noortgate & Onghena, 2008).

Secondly, the BC-SMD effect size is defined based on hierarchical linear models that assume Gaussian errors. The assumption of Gaussian errors might work reasonably well for typical academic outcomes, but can be a poor approximation for some outcomes in the form of frequency counts or percentage durations—particularly when the baseline level of the outcome is near zero. If the distributional assumptions of the model are violated, the variance component estimators may be biased (Declercq et al., 2018), leading to bias in the

BC-SMD effect size estimator and in its standard error. Further research is needed to understand the robustness of current models to non-Gaussian error distributions and to determine design-comparable effect sizes can be defined and estimated for designs where Gaussian hierarchical linear models are not appropriate.[4]

Third, the estimation methods that we have applied are based on approximations that may not hold with very small samples of cases. For the two-level models described in earlier work on the BC-SMD, simulation evidence indicated that the methods provide close to unbiased effect size estimates even with a very limited number of cases, although accurate estimation of standard errors required more cases (Pustejovsky et al., 2014). The sample sizes in the three applications that we have presented are relatively large in terms of the number of participants or groups. Further investigations of small-sample performance and minimum sample size guidelines are needed for the multi-level models that we have applied in these more complex forms of multiple baseline designs. It may also be fruitful to investigate other estimation strategies, such as using model selection algorithms (Li et al., 2021), Bayesian methods (Swaminathan et al., 2014), or between-series estimators (Ferron et al., 2014; Joo et al., 2021).

Fourth, the BC-SMD methods require multiple participants for estimating the between-participant variation that is used in the definition and estimation of BC-SMD effect sizes. A minimum of three participants was recommended for the treatment reversal design with replication across participants or multiple baseline designs, but more participants might be needed for more complex model specification (Valentine et al., 2016). The three-level BC-SMD approach applies to replicated multiple baseline across behaviors or settings with multiple participants, or clustered multiple baseline design that includes multiple participants, or multivariate multiple baseline design across multiple participants. Given the model complexity in these variations, more participants are preferred for more

[4] This poses a particular challenge for multivariate models involving outcomes that might have different error structures, such as when one outcome can be modeled using Gaussian error distributions but another outcome should be modeled using a Poisson distribution for frequency counts.

precise estimation of effect sizes.

Finally, although we have extended the BC-SMD methods in Pustejovsky et al. (2014) to more complex multiple baseline designs, there remains a need for methods that can be used with other types of SCDs, such as multi-element designs, alternating treatments designs, adapted alternating treatment designs, and repeated acquisition designs (Kirby et al., 2021), as well as to hybrid designs such as multiple baseline designs with embedded treatment reversals or alternating treatment phases. Recent work has examined inferential methods and graphical representations for alternating treatment designs (Manolov et al., 2021), but these methods focus on the individual case level and do not provide a summary effect size in the same metric as a group design. Extensions for such designs are a valuable direction for further work because the designs are commonly used in classroom setting to compare multiple treatments (Kazdin, 2011).

Conclusion

One of the major concerns over use of the BC-SMD for summarizing findings from single-case studies was that doing so leads to exclusion of evidence from experimentally valid designs, purely due to the technical requirements of the effect size estimation methods (Kratochwill et al., 2021; Maggin et al., 2021). The methods that we have described provide a way to estimate BC-SMD effect sizes for a broader range of single-case designs than was previously possible. These designs are well-aligned to how interventions are conducted in schools, and the replicated multiple baseline across behaviors design provides a more rigorous evaluation of an intervention by incorporating within-participant replication. The proposed extensions should mitigate concerns over the technical limitations of BC-SMDs at least partially, although the methods still require data from multiple participants. This is an inherent constraint required to achieve comparability between the effect sizes derived from SCDs and those from between-group designs. At the same time, estimation of BC-SMD effect sizes requires careful model building and can be

sensitive to the assumptions made, especially those regarding time trends and whether trends and treatment effects vary across cases. Furthermore, it requires the analyst to be specific about the form of the hypothetical between-group design for which the effect size is defined. The aspects of the methods arise inherently because the BC-SMD aims to provide a common metric with standardized mean differences from between-group designs.

It is worth recognizing that the BC-SMD effect size is just one metric, among many others, for quantifying the intervention effects from single-case studies. There will certainly be situations where the limitations of using between-case effect sizes outweigh their advantages. It is critical that researchers select an appropriate effect size based on the context and purpose of their work.

Data and Replication Materials

Raw data from all of the examples presented in the paper and R code for replicating all reported analyses are available on the Open Science Framework at https://osf.io/8eucf.

Acknowledgements

We gratefully acknowledge Diane and Brian Bryant for sharing student-level data from their clustered multiple baseline design (Bryant et al., 2018). We thank David Kaplan and Jee-Seon Kim for providing feedback on this work.

References

Babb, S., Raulston, T. J., McNaughton, D., Lee, J.-Y., & Weintraub, R. (2021). The Effects of Social Skill Interventions for Adolescents With Autism: A Meta-Analysis. *Remedial and Special Education*, *42*(5), 343–357. https://doi.org/10.1177/0741932520956362

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

Barker, J. B., Mellalieu, S. D., McCarthy, P. J., Jones, M. V., & Moran, A. (2013). A review of single-case research in sport psychology 1997–2012: Research trends and future directions. *Journal of Applied Sport Psychology*, *25*(1), 4–32. https://doi.org/10.1080/10413200.2012.709579

Barton, E. E., Ledford, J. R., Lane, J. D., Decker, J., Germansky, S. E., Hemmeter, M. L., & Kaiser, A. (2016). The iterative use of single case research designs to advance the science of EI/ECSE. *Topics in Early Childhood Special Education*, *36*(1), 4–14. https://doi.org/10.1177/0271121416630011

Bryant, D., Bryant, B., Sorelle-Miner, D., Falcomata, T., & Nozari, M. (2018). Tier 3 intensified intervention for second grade students with severe mathematics difficulties. *Archives of Psychology*, *2*(11). https://archivesofpsychology.org/index.php/aop/article/view/86

Busk, P. L., & Serlin, R. C. (1992). Meta-analysis for single-case research. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 187–212). Lawrence Erlbaum Associates, Inc.

Calder, S. D., Claessen, M., Ebbels, S., & Leitão, S. (2021). The efficacy of an explicit intervention approach to improve past tense marking for early school-age children with

developmental language disorder. *Journal of Speech, Language, and Hearing Research*, *64*(1), 91–104. https://doi.org/10.1044/2020_JSLHR-20-00132

Calder, S. D., Claessen, M., Ebbels, S., & Leitão, S. (2020). Explicit grammar intervention in young school-aged children with developmental language disorder: An efficacy study using single-case experimental design. *Language, Speech, and Hearing Services in Schools*, *51*(2), 298–316. https://doi.org/10.1044/2019_LSHSS-19-00060

Cook, B. G., Johnson, A. H., Maggin, D. M., Therrien, W. J., Barton, E. E., Lloyd, J. W., Reichow, B., Talbott, E., & Travers, J. C. (2021). Open science and single-case design desearch. *Remedial and Special Education*. https://doi.org/10.1177/0741932521996452

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

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

Gast, D. L., & Ledford, J. R. (2018). Research approaches in applied settings. In J. R. Ledford & D. L. Gast (Eds.), *Single case research methodology* (pp. 1–26). Routledge.

Gast, D. L., Lloyd, B. P., & Ledford, J. R. (2018). Multiple baseline and multiple probe designs. In *Single case research methodology* (pp. 239–281). Routledge.

Gilmour, A. R., Thompson, R., & Cullis, B. R. (1995). Average information REML: An efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, 1440–1450. https://doi.org/10.2307/2533274

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*, 1–15. https://doi.org/10.1080/2372966X.2020.1853486

Gurka, M. J. (2006). Selecting the best linear mixed model under REML. *The American Statistician*, *60*(1), 19–26. https://doi.org/10.1198/000313006X90396

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. (2008). What are effect sizes and why do we need them? *Child Development Perspectives*, *2*(3), 167–171. https://doi.org/10.1111/j.1750-8606.2008.00060.x

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

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

Joo, S.-H., Wang, Y., Ferron, J., Beretvas, S. N., Moeyaert, M., & Van Den Noortgate, W. (2021). Comparison of within- and between-series effect estimates in the meta-analysis of multiple baseline studies. *Journal of Educational and Behavioral Statistics*. https://doi.org/10.3102/10769986211035507

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

Kirby, M. S., Spencer, T. D., & Ferron, J. (2021). How to be RAD: Repeated acquisition design features that enhance internal and external validity. *Perspectives on Behavior Science*, *44*(2-3), 389–416. https://doi.org/10.1007/s40614-021-00301-2

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). *What works clearinghouse single-case design technical documentation version 1.0*.

https://ies.ed.gov/ncee/wwc/Docs/ReferenceResources/wwc_scd.pdf

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

Kratochwill, T. R., & Levin, J. R. (2014). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances* (pp. 53–89). American Psychological Association. https://doi.org/10.1037/a0017736

Ledford, J. R., Lambert, J., Pustejovsky, J. E., Zimmerman, K. N., & Barton, E. (2022). *Single case design research in special education: Next generation standards and considerations*. https://doi.org/10.31219/osf.io/e98nw

Li, H., Luo, W., Baek, E., Thompson, C. G., & Lam, K. H. (2021). Estimation and statistical inferences of variance components in the analysis of single-case experimental design using multilevel modeling. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01691-6

Losinski, M., Sanders, S., Katsiyannis, A., & Wiseman, N. (2017). A meta-analysis of interventions to improve the compliance of students with disabilities. *Education and Treatment of Children*, *40*(4), 435–463. https://doi.org/10.1353/etc.2017.0020

Maggin, D. M., Barton, E., Reichow, B., Lane, K., & Shogren, K. A. (2021). Commentary on the *What Works Clearinghouse Standards and Procedures Handbook* (v. 4.1) for the review of single-case research. *Remedial and Special Education*. https://doi.org/10.1177/07419325211051317

Manolov, R., Tanious, R., & Onghena, P. (2021). Quantitative techniques and graphical representations for interpreting results from alternating treatment design. *Perspectives on Behavior Science*. https://doi.org/10.1007/s40614-021-00289-9

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., Rindskopf, D., Onghena, P., & Van den Noortgate, W. (2017). Multilevel modeling of single-case data: A comparison of maximum likelihood and bayesian estimation. *Psychological Methods*, *22*(4), 760.

Moeyaert, M., Zimmerman, K. N., & Ledford, J. R. (2018). Single case research methodology: Applications in special education and behavioral sciences. In J. R. Ledford & D. L. Gast (Eds.), *Single case research methodology* (pp. 393–416). Routledge.

Odom, S. L., Barton, E. E., Reichow, B., Swaminathan, H., & Pustejovsky, J. E. (2018). Between-case standardized effect size analysis of single case designs: Examination of the two methods. *Research in Developmental Disabilities*, *79*, 88–96. https://doi.org/10.1016/j.ridd.2018.05.009

Parker, R. I., Hagan-Burke, S., & Vannest, K. (2007). Percentage of all non-overlapping data (PAND) an alternative to PND. *The Journal of Special Education*, *40*(4), 194–204. https://doi.org/10.1177/00224669070400040101

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

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2019). *nlme: Linear and nonlinear mixed effects models*. https://CRAN.R-project.org/package=nlme

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

Pustejovsky, J. E. (2019). Procedural sensitivities of effect sizes for single-case designs with directly observed behavioral outcome measures. *Psychological Methods*, *24*(2), 217. https://doi.org/10.1037/met0000179

Pustejovsky, J. E., Chen, M., & Hamilton, B. J. (2020a). *scdhlm: A web-based calculator*

for between-case standardized mean differences (0.5.2) [Web application]. https://jepusto.shinyapps.io/scdhlm

Pustejovsky, J. E., Chen, M., & Hamilton, B. J. (2020b). *scdhlm: Estimating hierarchical linear models for single-case designs* (0.5.0) [R package]. https://jepusto.github.io/scdhlm

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

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

Radley, K. C., Dart, E. H., Fischer, A. J., & Collins, T. A. (2020). Publication trends for single-case methodology in school psychology: A systematic review. *Psychology in the Schools*, *57*(5), 683–698. https://doi.org/10.1002/pits.22359

Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The quantitative synthesis of single-subject research: Methodology and validation. *Remedial and Special Education*, *8*(2), 24–33. https://doi.org/10.1177/074193258700800206

Shadish, W. R. (2014). Analysis and meta-analysis of single-case designs: An introduction. *Journal of School Psychology*, *52*(2), 109–122. https://doi.org/10.3758/s13428-011-0111-y

Shadish, W. R., Hedges, L. V., Horner, R. H., & Odom, S. L. (2015). *The Role of Between-Case Effect Size in Conducting, Interpreting, and Summarizing Single-Case Research* (NCER 2015-002; p. 109). National Center for Education Research, Institute of Education Sciences, U.S. Department of Education. https://eric.ed.gov/?id=ED562991

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.

Shin, M., Bryant, D., Powell, S. R., Jung, P.-G., Ok, M. W., & Hou, F. (2020). *A meta-analysis of single-case research on word-problem instruction for students with learning disabilities*. https://doi.org/10.31219/osf.io/qjtg6

Simonsen, B., Freeman, J., Myers, D., Dooley, K., Maddock, E., Kern, L., & Byun, S. (2020). The effects of targeted professional development on teachers' use of empirically supported classroom management practices. *Journal of Positive Behavior Interventions*, *22*(1), 3–14. https://doi.org/10.1177/1098300719859615

Skinner, C. H., Mccleary, D. F., Skolits, G. L., Poncy, B. C., & Cates, G. L. (2013). Emerging opportunities for school psychologists to enhance our remediation procedure evidence base as we apply response to intervention. *Psychology in the Schools*, *50*(3), 272–289. https://doi.org/10.1002/pits.21676

Snijders, T. A., & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. sage.

Swaminathan, H., Rogers, H. J., & Horner, R. H. (2014). An effect size measure and bayesian analysis of single-case designs. *Journal of School Psychology*, *52*(2), 213–230. https://doi.org/10.1016/j.jsp.2013.12.002

Tarlow, K. R. (2017). An improved rank correlation effect size statistic for single-case designs: Baseline corrected Tau. *Behavior Modification*, *41*(4), 427–467. https://doi.org/10.1177/0145445516676750

Taylor, J. A., Pigott, T., & Williams, R. (2021). Promoting knowledge accumulation about intervention effects: Exploring strategies for standardizing statistical approaches and effect size reporting. *Educational Researcher*, 72–80. https://doi.org/10.3102/0013189X211051319

Texas Education Agency/University of Texas System. (2009). *Texas early mathematics inventory—aim check.* Authors, Austin, TX.

Thiemann, K. S., & Goldstein, H. (2001). Social stories, written text cues, and video feedback: Effects on social communication of children with autism. *Journal of Applied*

Behavior Analysis, *34*(4), 425–446. https://doi.org/10.1901/jaba.2001.34-425

Valentine, J. C., Tanner-Smith, E. E., Pustejovsky, J. E., & Lau, T. (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

Van den Noortgate, W., & Onghena, P. (2008). A multilevel meta-analysis of single-subject experimental design studies. *Evidence-Based Communication Assessment and Intervention*, *2*(3), 142–151. https://doi.org/10.1080/17489530802505362

What Works Clearinghouse. (2020). *WWC procedures and standards handbook (version 4.1). Washington, DC: US department of education, institute of education sciences.* https://ies.ed.gov/ncee/wwc/Handbooks

Wright, M. R., Ellis, D. N., & Baxter, A. (2012). The effect of immediate or delayed video-based teacher self-evaluation on Head Start teachers' use of praise. *Journal of Research in Childhood Education*, *26*(2), 187–198. https://doi.org/10.1080/02568543.2012.657745