Design-Comparable Effect Sizes in Multiple Baseline Designs: A General Modeling Framework

James E. Pustejovsky

The University of Texas at Austin

Larry V. Hedges

Northwestern University

William R. Shadish

University of California, Merced

Authors' Note

The research presented in this article is based on the first author's Ph.D. dissertation, which was completed at Northwestern University under the supervision of Larry Hedges. The opinions expressed are those of the authors and do not represent views of the funders.

Funding

This research was supported in part by a grant from the Institute for Educational Sciences (R305D100046).

Authors

James E. Pustejovsky is an assistant professor in the Educational Psychology Department at the University of Texas at Austin, Department of Educational Psychology, 1 University Station D5800, Austin, TX 78712-1294. email: pusto@austin.utexas.edu. His interests include statistical methods for meta-analysis and statistical models and analytic methods for single-case research.

Larry V. Hedges is the Board of Trustees Professor of Statistics, Psychology, Education, and Social Policy and a Faculty Fellow of the Institute for Policy Research, Northwestern University, 2040 Sheridan Road, Evanston, IL 60208-4100; email: l-hedges@northwestern.edu. His interests include statistical methods for meta-analysis and design and analysis of social experiments.

William R. Shadish is Distinguished Professor at the University of California-Merced, School of Social Sciences, Humanities and Arts, 5200 North Lake Rd, Merced, CA 95343; e-mail: wshadish@ucmerced.edu. His interests include experimental and quasi-experimental designs, program evaluation, and meta-analysis.

Abstract

In single-case research, the multiple baseline design is a widely used approach for evaluating the effects of interventions on individuals. Multiple baseline designs involve repeated measurement of outcomes over time and the controlled introduction of a treatment at different times for different individuals. This article outlines a general approach for defining effect sizes in multiple baseline designs that are directly comparable to the standardized mean difference from a between-subjects randomized experiment. The target, design-comparable effect size parameter can be estimated using restricted maximum likelihood together with a small-sample correction analogous to Hedges' *g*. The approach is demonstrated using hierarchical linear models that include baseline time trends and treatment-by-time interactions. A simulation compares the performance of the proposed estimator to that of an alternative, and an application illustrates the model-fitting process.

Substantive keywords: Special Education, Counseling, School Psychology

Methodological keywords: single-subject research; effect size; meta-analysis; hierarchical linear model

Design-Comparable Effect Sizes in Multiple Baseline Designs: A General Modeling Framework

One of the central questions in any quantitative research synthesis is how to operationally define the effect size, the basic unit of analysis in a meta-analysis (Cooper, 2009). Cohen's *d* is a common choice for two-group comparative experiments with continuous outcomes. This effect size has the form of a ratio where the numerator represents a treatment effect (in the scale of the outcome variable) and the denominator represents the standard deviation of the control group (or of both the control and treatment groups, if these are assumed to be the same in the population). In simple experimental designs, these quantities are unambiguous. However, the choice of effect size is less simple in more complex designs, even if one considers only effect sizes of the same form as Cohen's *d*. In cluster-randomized experiments with two or more levels, several different *d*-type effect sizes may be considered, each of which standardizes the treatment effect by a different combination of variance components (Hedges, 2007, 2011). Similarly, longitudinal designs identify multiple components of variation, including within-individual variation across measurement occasions and variation across individuals; further, both the treatment effect and the variance components might change over the time period being studied.

The choice of effect size involves, implicitly or explicitly, an assumption about the comparability of results from different studies that may use various measurement instruments, inclusion criteria, or experimental designs. In synthesizing results from studies that use different types of designs, one would like to use an effect size that is on the same metric across all of them. Such an effect size, which we call *design-comparable*, is desirable so that contrasts between and averages across results from different types of designs are not confounded by

differences in scale. Without design-comparability, a collection of effect sizes will exhibit heterogeneity due only to differences in how they are calculated, obscuring substantive differences that may exist. Design-comparability thus allows the analyst to adjust for incidental aspects of study design and to focus instead on variation that is of scientific interest. This paper addresses the question of how to find and estimate design-comparable effect sizes for a certain type of experimental design, the multiple baseline design, which falls in the broader category of single case designs.

Single case designs (SCDs) are a class of experiments used to evaluate the effects of interventions or treatment programs. Two features characterize the class: (1) repeated measurement of an outcome over time on one or more individual cases and (2) identification of effects through comparison of each individual's outcomes in the presence or absence of treatment, across different points in time. SCDs are used frequently in fields such as applied behavior analysis (Bailey & Burch, 2002), special education research (Horner et al., 2005), school psychology (Busse, Kratochwill, & Elliott, 1995), and psychotherapy (Borckardt et al., 2008).

The multiple baseline design is the most common type of SCD (Shadish & Sullivan, 2011; Smith, 2012). The design involves one or more individual cases, on each of which an outcome is measured multiple times during a baseline phase. After several measurements, a treatment is introduced for one case (but not others) and outcome measurements are continued for all cases. Treatment is introduced for a second case at a subsequent time, a third after that, and so on, until all cases have received treatment. The treatment is taken to have an effect for a given case if the stable pattern of outcomes differs between baseline and treatment phases and if the change coincides with the introduction of treatment. The single case design literature

describes several types of multiple baseline designs, including multiple baselines across individuals, multiple baselines across settings, and multiple baselines within individuals across outcomes (Kazdin, 2011); in what follows, we restrict attention to the first of these.

Meta-analytic synthesis has long been considered as an approach for generalizing from single case studies (Center, Skiba, & Casey, 1985; Gorsuch, 1983; Scruggs, Mastropieri, & Casto, 1987), yet despite long-standing interest, there is still little consensus regarding how single-case studies should be synthesized. Many different effect size metrics have been proposed for SCDs, though nearly all are subject to serious criticisms (Beretvas & Chung, 2008; Wolery, Busick, Reichow, & Barton, 2010). Among their shortcomings, most of the proposed effect sizes are not directly comparable to those from between-subjects designs.

A design-comparable effect size for single-case designs was proposed by Hedges, Pustejovsky, and Shadish (HPS hereafter), who used a certain hierarchical linear model to define a *d*-type effect size index and explicitly demonstrate its equivalence to Cohen's *d* from a between-subjects randomized experiment. HPS proposed methods for estimating the effect size from data collected in treatment reversal designs (Hedges, Pustejovsky, & Shadish, 2012) and multiple baseline designs (Hedges, Pustejovsky, & Shadish, 2013). For both designs, the effect size estimators were designed to be expressible using closed-form algebra (albeit in formulas that are too extensive to describe here) and to be relatively insensitive to estimated nuisance parameters. However, HPS considered only a single data-generating model that relied on several restrictive assumptions, including that baselines are stable, lacking any trends; that the treatment effect can be modeled by a shift in the mean level of the outcome; and that the treatment effect is homogeneous across cases. The simple model considered by HPS (MB1, described below) led to a correspondingly simple operational definition for the design-comparable effect size, but the

strong assumptions entailed by the model limit the set of studies where the effect size described by HPS can be applied. In contrast to the posited model, multiple baseline studies frequently exhibit trends in both baseline and treatment phases and single case researchers often operate in contexts with high inter-case variability (Hersen, 1990). Furthermore, it is not apparent how the HPS estimation approach could be extended to more general models.

In this paper, we demonstrate how the general logic behind the HPS effect size can be applied to define and estimate similar *d*-type effect sizes under a variety of more flexible models for multiple baseline designs. Specifically, we focus on models that incorporate linear time trends (either homogeneous or varying across cases) and heterogeneous treatment effects, though the same general logic is applicable in principle to more complex models as well. These models entail treatment effects and variance components that may change over the time, both of which make effect size definition more complicated. We address these complications by operationalizing an effect size that is specific to both a fixed length of treatment and a fixed point in time. After describing an approach to modeling and effect size definition, we describe a method for estimating these effect sizes from multiple baseline designs. The proposed effect size estimator, based on a small-sample correction to the restricted maximum likelihood estimator, is more flexible and more readily extensible than the HPS estimator. A later section summarizes the results of several simulation studies that evaluate the small-sample performance of the proposed method (full simulation results are available in the supplementary materials). We then demonstrate the proposed approach with an application to data from a real multiple baseline design. The final section concludes by noting further extensions and outstanding questions.

Constructing design-comparable effect sizes

In this section, we describe how to construct design-comparable effect sizes for multiple baseline designs that allow direct comparisons to Cohen's *d* effect sizes from between-subjects designs. We first lay out notation and specify the design of a multiple baseline experiment. We then discuss our general modeling framework and method of defining a target, design-comparable effect size parameter.

Consider a set of m individuals and N measurement occasions; for purposes of exposition, we assume that the measurement occasions are equally spaced. Procedurally, a multiple baseline design across individuals entails the following. For case i, T_i measurements are made during a baseline phase (in the absence of treatment), where $0 \leq T_i \leq N$. The case then receives treatment, and $N - T_i$ further outcome measurements are made during the treatment phase. If no observations are missing, the data generated from this design consist of $m \times N$ measurements, which we denote Y_{ij} for $i = 1,\ldots,m$ and $j = 1, \ldots, N$. The key feature of the multiple baseline design is that not all cases have the same point of treatment introduction, so $T_h \neq T_i$ for one or more pairs $h, i = 1, ..., m$.

Our approach requires specification of an appropriate statistical model for the multiple baseline data. In order for the statistical model to be general enough that it also encompasses a between-subjects experiment, the model must be causally interpretable at the level of the individual case, must specify a treatment assignment mechanism, and must capture variation both within and between cases. By causally interpretable, we mean that the model clearly describes not just the observed outcome data, but also the outcomes that would be observed under variations in how treatment is assigned. These are often called potential outcomes (Holland, 1986). An interpretable model entails writing the outcome for case i at time j as a function of the treatment assignment time for that case, that is, as $Y_{ij}(T_i)$. In a multiple baseline

design, the treatment phase for case *i* might begin after any given measurement occasion, so that T_i takes values from 0, ..., *N*, where $T_i = 0$ if the case is assigned to treatment just prior to the first outcome measurement.

A causally interpretable model describes many data points, only some of which are observed in a given design. It will often be reasonable to assume that outcomes do not depend on future treatment assignment times—that is, the outcome on Monday does not depend on whether a case will receive the treatment on the next Thursday versus on the next Friday. Even with this restriction, a causally interpretable model for case *i* describes many potential outcome values, only a few of which are observed in a given design: consider that $Y_{ij}(T_i)$ for $j = 1, ..., N$ and $T_i = 0, ..., j$ involves a total of $N \times (N + 1) / 2$ unique values, only *N* of which can be observed. We specify a simple, causally interpretable case-level model in the following section.

Given a causally interpretable model, one must also consider the relationship between the potential outcomes and the data actually observed in the experiment. In all of the models considered below, we assume that the baseline phase lengths $T_1,\ldots, T_m$ are assigned independently of the potential outcomes, as would be the case with random assignment of cases to intervention start times. This excludes certain mechanisms such as introducing the treatment first to the case with the highest baseline outcomes, second to the case with the next-highest baseline, etc. Such non-random assignment mechanisms introduce substantial complications that will need to be addressed in future work.

Finally, for the statistical model to be sufficiently general, it must also make explicit assumptions about variation across cases. This can be done quite naturally using the framework of hierarchical models, an approach that has recently begun to receive increased attention for the analysis and meta-analysis of single case designs.

A sufficiently general statistical model allows one to specify precisely a design-comparable effect size parameter. To describe this effect size, it is helpful to consider a hypothetical, between-subjects experiment, in which treatment assignment begins at a fixed point in time, a set schedule of treatment follows, and outcomes are measured at a fixed, later point in time. Though this is clearly a stylization of how any actual experiment occurs, it is nonetheless useful in that it makes apparent how a design-comparable effect size can depend on the specified time-points.

Imagine a between-subjects experiment in which treatment begins immediately following a certain measurement occasion *A* (the implementation time) and in which outcomes are measured on all units at a later time *B* (the follow-up time or end-point). For multiple baseline designs, these design parameters are all that is needed to describe a design-comparable effect size. Once the time-points are specified, a design-comparable *d*-type effect size is given by

$$\delta_{AB} = \frac{\mathrm{E}\left(Y_{iB}(A)\right) - \mathrm{E}\left(Y_{iB}(N)\right)}{\sqrt{\mathrm{Var}\left(Y_{iB}(N)\right)}}. \qquad (1)$$

This parameter represents the difference between the average outcome if treatment is introduced at time *A* and the average outcome if treatment is never introduced, divided by the standard deviation of the outcome if treatment is never introduced, where all outcomes are measured at a fixed time *B*. It is the same parameter as the estimand of Cohen's *d* from the hypothetical between-subjects experiment with implementation time *A* and follow-up time *B*. Note that the definition of δ_{AB} relies on having a causally interpretable model for the data, in order for the quantities $Y_{iB}(A)$ and $Y_{iB}(N)$ to be defined. In each model considered below, we will re-express the general parameter defined by (1) in terms of model components. As will be seen, in some models the effect size may depend only on *B* - *A*, only on *B*, or on both *A* and *B*.

Having outlined our general approach to constructing comparable effect sizes for multiple baseline designs, we now turn to the details. The next section describes several model specifications for multiple baseline designs and demonstrates how the general approach is applied in each case to find the effect size parameter. The following section discusses estimation of the effect size parameter.

Models and effect size parameters

This section presents a catalog of model specifications for multiple baseline designs, along with the corresponding effect size parameters. We write the models using a two-level formulation similar to that used by Singer and Willett (2003), where level one (the within-case level) describes a regression model for the i^{th} individual and level two (the between-case level) describes how the within-case regression coefficients vary across cases. As will be seen, the specifications under consideration can all be written using the same within-case assumptions; only the between-case assumptions change.

The within-case assumptions

Our general framework requires a within-case model that is causally interpretable. We therefore specify a structural model where the outcome Y_{ij} is a function of treatment assignment time T_i, as follows:

$$Y_{ij}(T_i) = \beta_{0i} + \beta_{1i}1(j > T_i) + \beta_{2i}(j - C) + \beta_{3i}\left((j - T_i) \times 1(j > T_i)\right) + \epsilon_{ij}, \qquad (2)$$

Here and following, $1(j > T_i)$ is an indicator variable equal to 0 for $j \leq T_i$ and to 1 for $j > T_i$. Equation (2) is a piece-wise linear regression model, a very common specification for analysis of multiple baselines (Center et al., 1985; Gottman, 1981; Huitema & McKean, 2000). In this model, time is taken to be equivalent to measurement occasion and is centered at the constant C.

To interpret the coefficients, it is helpful to first consider the form of the regression if case *i* were to remain in the baseline phase for the entire study, so that $T_i = N$. The model then reduces to $Y_{ij}(N) = \beta_{0i} + \beta_{2i}\left(j - C\right) + \epsilon_{ij}$, and it can be seen that β_{0i} represents the average level of the outcome at time $j = C$ in the absence of treatment and β_{2i} represents the linear change in the outcome per measurement occasion, also in the absence of treatment. For arbitrary T_i, β_{1i} and β_{3i} describe the effect of the treatment on case *i*. Specifically, β_{1i} represents the immediate change in the level of the outcome due to introducing the treatment, while β_{3i} represents additional change in the outcome per measurement occasion that is due to the treatment. If the treatment began just after time *A*, then the individual treatment effect for case *i* at time *B* would be $\beta_{1i} + \beta_{3i}$ $(B - A)$.

It remains to describe assumptions about the error term ε_{ij}. Because measurements on each case are taken repeatedly over time, the assumption that the errors are independent is considered implausible. In the literature on statistical analysis of single-case designs, the most common assumption is that the errors are auto-correlated and follow a stationary, first-order auto-regressive process. We follow convention by assuming that the errors have expectation zero, variance σ^2, and first-order autocorrelation ϕ. The last assumption implies that $\mathrm{Cov}\left(\epsilon_{ij}, \epsilon_{ik}\right) = \phi^{|k-j|}\sigma^2$. Further, all errors are assumed to be independent across cases, so $\mathrm{Cov}\left(\epsilon_{hj}, \epsilon_{ik}\right) = 0$ if $h \neq i$. Of course, many other assumptions regarding the error structure could be considered (see, e.g., Pinheiro & Bates, 2000, sec. 5.3). Having described the within-case assumptions, the remainder of this section examines several between-case specifications that make different assumptions about variation across individuals.

Model MB1: Varying intercepts, fixed treatment effect, no trends

HPS considered perhaps the simplest possible model for multiple baseline data, assuming that baseline outcomes are stable (lacking trend) and that the treatment causes a shift in the level

of the outcome that is constant across individuals. Using the within-case model (2), their model is equivalent to assuming that

$$\beta_{0i} = \gamma_{00} + \eta_{0i}, \quad \beta_{1i} = \gamma_{10}, \quad \beta_{2i} = 0, \quad \beta_{3i} = 0, \tag{3}$$

where η_{0i} is normally distributed with mean zero and variance τ_0^2. Here, γ_{00} is the average level of the outcome across individuals in the absence of treatment, $\eta_{0i} = \beta_{0i} - \gamma_{00}$ is the deviation from this average level for case *i*, and γ_{10} is the treatment effect, assumed to be constant across individuals.[1] The coefficients for the time trends β_{2i} and the time-by-treatment interactions β_{3i} are both assumed to be zero. The effect size parameter for MB1 follows by evaluating Expression (1) using (2) and (3):

$$\delta_{AB} = \frac{\gamma_{10}}{\sqrt{\tau_0^2 + \sigma^2}}. \tag{4}$$

In a later section, we will describe a method for estimating δ_{AB} that differs from the one proposed by HPS.

Model MB2: Varying intercepts, varying treatment effects, no trends

MB1 makes the restrictive assumption that the treatment effect γ_{10} is constant across cases. This assumption can be relaxed by allowing the treatment effect to vary across individuals, while retaining the assumptions regarding the stability of baseline and treatment phases (cf. Ferron, Bell, Hess, Rendina-Gobioff, & Hibbard, 2009). The between-case specification becomes:

$$\beta_{0i} = \gamma_{00} + \eta_{0i}, \quad \beta_{1i} = \gamma_{10} + \eta_{1i}, \quad \beta_{2i} = 0, \quad \beta_{3i} = 0, \tag{5}$$

where (η_{0i}, η_{1i}) is multi-variate normally distributed, with mean (0, 0) and covariance matrix

$$\mathrm{T} = \begin{bmatrix} \tau_0^2 & \tau_{10} \\ \tau_{10} & \tau_1^2 \end{bmatrix}.$$

Because the effect size is scaled by the standard deviation of the outcome in the absence of treatment, allowing β_{1i} to vary randomly does not alter the parameter of interest; instead, its form is the same as the effect size under MB1, as given in (4). Though the parameter is identical, the assumption that the treatment effect is not constant across individuals has implications for how the parameter is estimated.

Model MB3: Varying intercepts, fixed treatment effect, fixed trends

Multiple baseline data often exhibit trends during the baseline phase, treatment phase, or both. A slightly less restrictive model than MB1 would allow for trends in both the baseline and treatment phase, but assume that those trends are common across individuals. Along with the assumption that the treatment has a constant effect across cases, the between-case model becomes:

$$\beta_{0i} = \gamma_{00} + \eta_{0i}, \quad \beta_{1i} = \gamma_{10}, \quad \beta_{2i} = \gamma_{20}, \quad \beta_{3i} = \gamma_{30}, \tag{6}$$

where $\eta_{0i} \sim N(0, \tau_0^2)$, as in MB1. The parameters γ_{00} and η_{0i} have the same interpretation as in MB1, but now γ_{10} represents the immediate change in the outcome after introducing treatment, γ_{20} represents the change in the outcome per measurement occasion in the absence of treatment, and γ_{30} represents additional change in the outcome per measurement occasion due to introducing treatment. All of γ_{10}, $\gamma_{20,}$ and γ_{30} are assumed to be constant across individuals. Under the assumptions of MB3, the effect size parameter is

$$\delta_{AB} = \frac{\gamma_{10} + \gamma_{30}(B - A)}{\sqrt{\tau_0^2 + \sigma^2}}. \tag{7}$$

Here, the effect size parameter depends on the difference $B - A$, the length of time between treatment introduction and outcome measurement. However, if $B - A$ is held constant, the

parameter does not depend on the choice of *B* alone, because the variance is constant across measurement occasions, regardless of the pattern of treatment assignments.

Model MB4: Varying intercepts, fixed treatment effects, varying trends

In some multiple baseline studies, the assumption in MB3 that the baseline slopes are constant across cases may itself be overly restrictive. To relax this assumption, let

$$\beta_{0i} = \gamma_{00} + \eta_{0i}, \quad \beta_{1i} = \gamma_{10}, \quad \beta_{2i} = \gamma_{20} + \eta_{2i}, \quad \beta_{3i} = \gamma_{30}, \tag{8}$$

where (η_{0i}, η_{2i}) is multi-variate normally distributed, with mean (0, 0) and covariance matrix

$$\mathrm{T} = \begin{bmatrix} \tau_0^2 & \tau_{20} \\ \tau_{20} & \tau_2^2 \end{bmatrix}.$$

Under this model, the design-comparable effect size is

$$\delta_{AB} = \frac{\gamma_{10} + \gamma_{30}(B - A)}{\sqrt{\tau_0^{\,2} + (B - C)^2 \tau_2^2 + 2(B - C)\tau_{20} + \sigma^2}}. \tag{9}$$

This parameter depends on the choice of both *A* and *B*, rather than just their difference, because the control-group variance changes over time.

Model MB5: Varying intercepts, varying trends, varying treatment-by-time interaction

To illustrate how hierarchical models can be tailored to specific contexts, we consider one further model here, to be applied in the example presented later. MB5 elaborates on MB4 by assuming that the treatment-by-time trend interaction varies across cases, while holding fixed the immediate treatment effect (the β_{1i}). The within-case assumptions for this model are identical to those in the previous models, as given in (2). The between-case assumptions are as follows:

$$\beta_{0i} = \gamma_{00} + \eta_{0i}, \quad \beta_{1i} = \gamma_{10}, \quad \beta_{2i} = \gamma_{20} + \eta_{2i}, \quad \beta_{3i} = \gamma_{30} + \eta_{3i},$$

where (η_{0i}, η_{2i}, η_{3i}) is multivariate normally distributed, with mean (0, 0, 0) and covariance matrix

$$\mathrm{T} = \begin{bmatrix} \tau_0^2 & \tau_{20} & \tau_{30} \\ \tau_{20} & \tau_3^2 & \tau_{32} \\ \tau_{30} & \tau_{32} & \tau_3^2 \end{bmatrix}.$$

In other words, in the absence of treatment, cases vary in their average levels of the outcome and in their rates of change; furthermore, the treatment has variable effects, altering the rate of change by more for some cases and less for others. Finally, note that the design-comparable effect size is equivalent to that given in Equation (9), because MB5 differs from MB4 only in the variability of the treatment effect, rather than baseline variability.

Further models

The five models presented above are far from an exhaustive list of the possible specifications. For example, one might assume in MB3 or MB4 that the time trend is constant across phases, so that $\beta_{3i} = 0$. Alternately, any of the models might be extended through the addition of polynomial time trends. In principle, one could even specify a model in which any or all of the within-case regression coefficients vary randomly, though in practice the number of randomly varying coefficients will need to be tempered by the number of cases m being modeled. Given an extended model, the effect size parameter can be found by evaluating Expression (1) under the structural assumptions of that model.

Estimation methods

We propose to estimate the effect size parameter by first estimating the model components via restricted maximum likelihood (REML), then making a small-sample correction. In earlier work, HPS developed specialized estimation approaches that are relatively insensitive to the method used for estimating nuisance parameters that are not of direct interest, and which provide close-to-unbiased estimates of the effect size parameters. However, the approaches rely on restrictive assumptions about the model and data under consideration, making them difficult

to generalize for more complex models. By comparison, REML estimation is extensible and accessible, in that it can accommodate many different models and is available in many common software packages. Because REML estimation is conventional, we focus on the details of effect size estimation rather than estimation of the basic model parameters. To begin, we formulate the target effect size parameter in general terms.

The hierarchical models discussed above have two sets of parameters: fixed effects, which we denote by γ's, and variance components, which include the auto-correlation coefficient ϕ, the within-case variance σ^2, and the random effect covariances denoted by τ's. Denote the ($p \times 1$) vector of fixed effects by $\boldsymbol{\gamma} = (\gamma_{00}, \ldots, \gamma_{(p-1)0})^T$ and the ($r \times 1$) vector that collects all of the variance component parameters by $\boldsymbol{\omega}$. Any of the effect sizes described in the previous section can then be expressed as the ratio of a linear combination of the fixed effects to the square root of a linear combination of the variance parameters. Given an appropriate parameterization of the variance components,

$$\delta_{AB} = \frac{\mathbf{p}^T \boldsymbol{\gamma}}{\sqrt{\mathbf{r}^T \boldsymbol{\omega}}}, \tag{10}$$

for suitably chosen ($p \times 1$) vector $\mathbf{p}$ and ($r \times 1$) vector $\mathbf{r}$. For example, in MB1, take $\boldsymbol{\gamma} = (\gamma_{00}, \gamma_{10})^T$ and $\boldsymbol{\omega} = \left(\sigma^2, \phi, \tau_0^2\right)^T$. Setting $\mathbf{p} = (0, 1)^T$ and $\mathbf{r} = (1, 0, 1)^T$ makes expression (10) equivalent to (4), the effect size defined for MB1. In MB4, take $\boldsymbol{\gamma} = (\gamma_{00}, \gamma_{10}, \gamma_{20}, \gamma_{30})^T$ and $\boldsymbol{\omega} = \left(\sigma^2, \phi, \tau_0^2, \tau_2^2, \tau_{20}\right)^T$. For a given choice of A, B, and centering point C, setting $\mathbf{p} = (0, 1, 0, B - A)^T$ and $\mathbf{r} = (1, 0, 1, [B - C]^2, 2[B - C])^T$ makes expression (10) equivalent to (9), the effect size defined for MB4.

Many software packages provide implementations of restricted maximum likelihood estimation for hierarchical linear models with auto-correlated residuals, including the `nlme`

package in R (Pinheiro, Bates, DebRoy, & Sarkar, 2012), the `xtmixed` command in Stata (StataCorp, 2011), `PROC MIXED` in SAS (SAS Institute Inc., 2008), and SPSS version 11.0 and following. All of these software packages provide REML estimates of both the fixed effect and variance component parameters, as well as an approximate covariance matrix for the fixed effects. All of the packages also provide some form of approximate covariance matrix of the variance components, though they differ in how this covariance matrix is estimated. Let $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\omega}}$ denote the REML estimates of the fixed effects and variance components, let the ($p \times p$) matrix $\mathbf{C}(\hat{\boldsymbol{\gamma}})$ denote the estimated covariance matrix of fixed effects, and let the ($r \times r$) matrix $\mathbf{C}(\hat{\boldsymbol{\omega}})$ denote the estimated covariance matrix of variance components.

An initial estimate of the effect size can be formed by substituting the REML estimates $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\omega}}$ in place of the corresponding parameters in (10), letting

$$\hat{\delta}_{AB} = \frac{\mathbf{p}^T \hat{\boldsymbol{\gamma}}}{\sqrt{\mathbf{r}^T \hat{\boldsymbol{\omega}}}}. \tag{11}$$

This unadjusted effect size estimate is approximately unbiased if the number of cases m is sufficiently large. However, for sample sizes typically found in multiple baseline designs, $\hat{\delta}_{AB}$ may nonetheless exhibit substantial bias even though $\hat{\boldsymbol{\gamma}}$ is exactly unbiased and $\hat{\boldsymbol{\omega}}$ is approximately unbiased.

The bias of $\hat{\delta}_{AB}$ is analogous to the small-sample bias of the Cohen's d statistic from a between-subjects experiment, which can be corrected using methods described by Hedges (1981); the corrected effect size estimate is sometimes referred to as Hedges' g.

The exact distribution theory used in Hedges' g statistic is not available for the present problem, due to the presence of additional nuisance parameters among the variance components. Still, one

can approximate the sampling distribution of $\hat{\delta}_{AB}$ with a Student-*t* distribution, thereby obtaining an approximate small-sample bias correction and an approximate expression for the variance of the effect size estimate.

From a theorem given in Hedges (2007), it follows that the distribution of $\hat{\delta}_{AB}$ can be approximated by a constant κ times a non-central *t* distribution with *v* degrees of freedom, where

$$\kappa = \sqrt{\frac{\mathbf{p}^T \mathbf{C}(\hat{\boldsymbol{\gamma}})\mathbf{p}}{\mathbf{r}^T \hat{\boldsymbol{\omega}}}} \tag{12}$$

and

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

It follows further that a bias-corrected effect size estimator is given by

$$g_{AB} = J(\nu) \times \hat{\delta}_{AB}, \tag{14}$$

where J(x) = 1 – 3 / (4x – 1). We will call g_{AB} the c-REML effect size estimator. Again approximating the distribution of g_{AB} by a non-central *t* distribution, an estimator for the approximate variance is given by

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

In a large-sample setting, a (1 – α) confidence interval (CI) for δ_{AB} could be constructed simply by taking the point estimate plus or minus the product of its standard error and the appropriate standard normal critical value. However, given the small number of cases typically found in single-case studies, we consider two other alternatives. The first is to construct a symmetric CI using a *t* distribution rather than the standard normal reference distribution:

$$g_{AB} \pm V_g \times t_{\alpha/2,\nu}, \tag{16}$$

where $t_{\alpha/2,\upsilon}$ is the $\alpha/2$ critical value from a central t distribution with υ degrees of freedom. Alternately, one could again appeal to the non-central t approximation to the distribution of $\hat{\delta}_{AB}$. Cumming and Finch (2001) described the construction of CIs for standardized mean differences based on a non-central t distribution. Algina and Keselman (2003) found that such CIs have superior coverage to symmetric CIs in both two-sample and repeated measures designs. Let $F_t\left(x\middle|\nu,\theta\right)$ be the cumulative distribution function of a non-central t random variable with ν degrees of freedom and non-centrality parameter θ. A $(1-\alpha)$ CI is then given by $[d_L, d_U]$, where the bounds are defined as the solutions to

$$\begin{aligned} F_t\left(\hat{\delta}_{AB}/\kappa \,\middle|\, \nu, d_L/\kappa\right) &= 1-\alpha/2 \\ F_t\left(\hat{\delta}_{AB}/\kappa \,\middle|\, \nu, d_U/\kappa\right) &= \alpha/2. \end{aligned} \tag{17}$$

We compare the small-sample performance of both of these CIs in the next section.

Small-sample performance

The estimation methods that we have proposed involve approximating the distribution of the REML effect size $\hat{\delta}_{AB}$ by a non-central t distribution. The small-sample performance of the c-REML estimator depends on the quality of this approximation, which may in turn depend on the particular data-generating model and the design of the study. In this section, we briefly summarize the results of three simulation studies examining the operating characteristics of the estimator under varying designs and data-generating models.

REML estimation methods are applicable in principle to a vast range of models and designs, but the scope of the simulations is necessarily much more limited. We address two main questions. First, how does the c-REML estimator compare to the effect size estimator proposed by HPS for model MB1? Under this simple model, the HPS estimator is nearly unbiased even in

designs with small numbers of cases and relatively few measurement occasions; for the c-REML estimator to be a viable alternative, it should have comparably small biases. Second, how does the proposed c-REML estimator perform under more complex models for multiple baseline design, where the HPS estimator is not immediately applicable? To address the latter question, we simulated the c-REML estimator under models MB2, which has treatment effects that vary across cases, and MB4, which has baseline trends that vary across cases.

In each simulation, we examined four properties of the proposed estimators. First, our foremost concern is with the bias of the point estimator. For purposes of meta-analysis, the estimator must be close to unbiased across the entire parameter space, because any remaining bias will be propagated through the meta-analysis. Second, we examined the precision of the point estimator, as measured by mean squared error (MSE). Third, we examined the relative bias of the proposed variance estimator. For purposes of meta-analysis, variance estimates are needed in order to assign appropriate weights when estimating grand average effects and meta-regressions and in order to estimate residual heterogeneity in random-effects models. Finally, we considered the actual coverage rates of the symmetric and non-central CIs described in the previous section, using conventional 95% coverage levels. Although CI coverage does not bear directly on the suitability of an effect size estimator for meta-analysis, a CI for the standardized mean difference is useful for characterizing the results of a single study.

A full description of the simulation designs, procedures, and results, including the tables and figures referenced in the remainder of this section, can be found in the supplementary materials.

Comparison of adjusted REML estimators and HPS estimator for MB1

The first simulation study compared the performance of the c- REML estimator to that of the estimator proposed by HPS under data-generating model MB1, using bias and root mean-squared error as the criteria. The simulation followed closely the design described in Hedges et al. (Hedges et al., 2013). We used a multiple baseline design in which treatment assignment times are spread as evenly as possible across the range of measurement occasions while maintaining at least 3 measurement occasions within each phase. We looked at two series lengths (N = 8 or N = 16) and varied the number of cases from m = 3 to m = 6; both of these dimensions are consistent with the sample sizes observed in empirical single-case research (Shadish & Sullivan, 2011). We varied the parameters of MB1 over the preponderance of the parameter space, using within-case reliability levels ranging from 0.0 to 0.8 and auto-correlation levels ranging from -0.7 to 0.7. This range of auto-correlations is consistent with the range of estimates from recent empirical single-case studies (Shadish & Sullivan, 2011), although we suspect that a narrower range may be more appropriate, given that part of the variability in auto-correlation estimates is due to measurement error (Shadish, Rindskopf, Hedges, & Sullivan, 2013).

The simulation results indicated the c-REML estimator has biases that are quite small and comparable to those of the HPS estimator, even at the smallest sample size considered. The bias of g_{AB} is no greater than 4.3% in absolute magnitude when estimated based on m = 3 cases; with m = 4 or m = 6 cases, the maximum absolute bias is 2.4% or 0.9%, respectively (Figure S1). Furthermore, the c-REML estimator has slightly smaller mean-squared error than the HPS estimator, on average and across sample sizes (Table S2).

Turning to variance estimation, we found that the c-REML variance estimator has smaller relative bias than the HPS variance estimator, and is less sensitive to the within-case reliability and auto-correlation (Figure S2). Averaging across the parameter levels, the variance estimator

for g_{AB} tends to over-estimate the true variance by 16% when $m = 3$ and $N = 8$, but by just 3% when $m = 4$ and $N = 8$. Together with the small biases of the point estimator, these results suggest that the c-REML estimator is a reasonable alternative to the HPS method for estimating effect sizes based on MB1.

Finally, we found that the symmetric CI tends to have greater than nominal coverage while the non-central CI had less than nominal coverage (Figure S3). Across the parameter levels and series lengths, the symmetric CI had average coverage ranging from 96.3% when $m = 3$ to 95.6% when $m = 6$, while the non-central CI had average coverage ranging from 91.9% when $m = 3$ to 93.5% when $m = 6$. When $m = 6$, the coverage of the symmetric CI ranged from 94.4% to 97.2% across the levels of the nuisance parameters and series length, while the coverage of the non-central CI was more variable, ranging from 91.1% to 95.6%. Overall, the symmetric CI is preferable under this model because its coverage is closer to nominal and errs conservatively towards over-coverage.

Performance of the adjusted REML estimator for MB2

The second simulation examined the performance of the c-REML estimator under data-generating model MB2, which allows treatment effects to vary across cases. The simulation design was identical to the previous simulation, but for the inclusion of an additional factor capturing treatment effect heterogeneity. The larger dimension of this model's parameter space prohibited as comprehensive an exploration of the parameter space. In order to moderate dimensionality, we limited the treatment effect heterogeneity factor to just two levels (moderate and high) and did not include a factor for the correlation between case-specific baseline levels and case-specific treatment effects.

Results indicated that the bias of the c-REML estimator remains small even when the number of cases is small (Figure S5). Across all combinations of parameters and series lengths, the absolute bias of g_{AB} is less than 7.3% when $m = 3$, less than 4.9% when $m = 4$, and less than 2.9% when $m = 5$. In contrast, the bias of the unadjusted REML estimator ($\hat{\delta}_{AB}$) is substantial: as much as 36.2% when $m = 3$, 23.7% when $m = 4$, and 17.6% when $m = 5$. Under MB2, the c-REML estimator appears to have biases small enough to warrant use in meta-analysis, particularly for studies with at least four cases.

The precision of the estimator is somewhat reduced under MB2 relative to MB1, due to both the increased variability of the treatment effects and the need to estimate additional variance components (Table S4). For instance, across levels of the nuisance parameters, the average MSE is 0.198 when $m = 4$ and $N = 8$, which is 28% larger than under MB1; the average MSE is 0.120 when $m = 6$ and $N = 8$, which is 30% larger than under MB1. As a point of comparison, a Hedges' g effect size estimate from a balanced, two-group experiment with 24 participants has MSE of 0.203 when the population effect size is one.

The variance estimator based on a non-central *t* approximation tends to have positive relative bias under MB2 (Figure S6). The over-estimation of variance is substantial when the number of cases is small, averaging 43% when $m = 3$ and $N = 16$. Even at the largest sample size, the variance of the effect size is moderately over-estimated, with average relative bias of 14% when $m = 6$ and $N = 16$. If the approximate variance estimator were used to apportion weights in the context of a fixed- or random-effects meta-analysis, the weight assigned to g_{AB} would tend to be understated relative to the theoretically most efficient weight. This would effectively down-weight evidence from single-case designs with very small numbers of cases, in a meta-analysis that included evidence from multiple types of designs. Though more accurate

variance estimates would certainly be preferable, use of the proposed variance approximation is conservative and thus not unreasonable for use in meta-analytic contexts, particularly for studies with larger numbers of cases.

The performance of the symmetric and non-central CIs follows a pattern similar to the previous simulation (Figure S7). Averaging across the parameter levels and series lengths, the symmetric CI had greater than nominal coverage, ranging from 97.8% when $m = 3$ to 96.4% when $m = 6$. When $m = 6$, the coverage rates varied from 93.9% to 97.6% across the levels of the auto-correlation, within-case reliability, and series length. Over-coverage is more extreme than under model MB1, likely due to over-estimation of the effect size variance. The non-central CI had less than nominal coverage, ranging from 93.3% when $m = 3$ to 94.2% when $m = 6$; at the largest sample size considered, the non-central CI had coverage rates ranging from 92.0% to 96.3%). As previously, the symmetric CI is preferred on the basis of conservatism.

Performance of the adjusted REML estimator for MB4

The final simulation study examined the performance of the c-REML estimator under data-generating model MB4, which allows for baseline trends that vary across cases and a change in trend due to treatment (constant across cases). Compared to MB1, MB4 includes additional parameters in both the mean specification and the variance components. We used a simulation design that was parallel to the design for MB2, but with two modifications. First, we replaced the factor capturing treatment effect heterogeneity with a factor capturing baseline trend heterogeneity. Second, we also included designs with a larger number of cases ($m = 9$ and $m = 12$). Further details can be found in the supplementary materials.

Similar to the results for MB2, the bias of the c-REML effect size estimator under MB4 is surprisingly small (Figure S8). When $m = 3$, the absolute bias of g_{AB} is less than 5.8% for the

shorter series length of $N = 8$, though it is as large as 12.8% for $N = 16$. When $m = 4$, the bias is never more than 2.7%, and the average bias across the parameter space is only -0.8%. As with MB2, the c-REML estimator is substantially less biased than the unadjusted estimator. Even under this more complex model, the bias is small enough that the point estimator should be considered suitable for use in meta-analysis, particularly for studies with at least four cases.

However, the effect size estimator is also very imprecise when the number of cases is small (Table S6). Across levels of the nuisance parameters, the average MSE is 0.444 when $m = 4$ and $N = 8$, which is nearly three times larger than under MB1. When $m = 6$ and $N = 8$, the average MSE is 0.255, which is 177% larger than under MB1 and comparable to the variance of Hedges' g statistic from a balanced, two-group experiment with 20 participants when the population effect size is one.

The proposed variance estimator has substantial relative biases when the number of cases is small (Figure S9). Unlike the previous simulation, the variance of the effect size tends to be under-estimated in model MB4. The underestimation is more extreme when $N = 16$; for this longer series length, the variance is underestimated by an average of 19% when $m = 4$, by 13% when $m = 6$, and by 5% when $m = 12$; the extent of underestimation also varies considerably across the parameter space. In model MB4, the variance approximation depends on a multiple of the treatment-by-trend interaction γ_{30}, and so may be particularly sensitive to understatement of the variance of the fixed effects. In summary, the proposed variance estimator requires a larger number of cases (or only a short extrapolation of the treatment-by-trend interaction) to perform adequately. Better estimators are needed if the study includes fewer than nine cases, because use of the proposed variance approximation for determining fixed- or random-effects meta-analytic

weights will tend to be anti-conservative, assigning too much weight to effect size estimates from model MB4.

The coverage rates of the symmetric and non-central CIs differed somewhat from the previous simulations. While the symmetric CI had greater than nominal coverage for the shorter series length of $N = 8$, the average coverage rate was less than nominal for the longer series length of $N = 16$ (Figure S10). The coverage of the non-central CI was less than nominal, with average coverage ranging from 88% when $m = 3$ to 92% when $m = 6$ to 93% when $m = 9$. When $m = 9$, the coverage rates of the central CI varied from 92.6% to 96.7% across the levels of the auto-correlation, within-case reliability, and series length. As previously, the symmetric CI should be preferred on the basis of conservatism, but should be taken as only a rough approximation when the number of cases is less than six. For both types of intervals, improved estimates of the fixed effects covariance matrix would likely improve coverage rates.

Application

We now present an application of the proposed effect size parameter and estimation methods, focusing in particular on the process of model fitting and comparison. Schutte, Malouff, and Brown (2008) evaluated the effect of an emotion-focused therapy program for adults with prolonged fatigue using a multiple baseline across individuals. The design included 13 adults who met clinical criteria for prolonged fatigue. Thus, this study is large compared to many multiple baseline designs, yet still small by statistical standards. Cases were measured weekly for 2, 5, or 8 weeks in the baseline phase and between 1 and 7 further weeks in the treatment phase, with a maximum of 15 consecutive measurements in all. For each case and measurement occasion, fatigue severity was measured using a self-reported scale that ranged from 1 to 63. Data for participant 4 are excluded from our analysis because nearly all of these

measurements are at the upper extreme of the scale, making the assumption of normally distributed errors implausible for this case. Data for the remaining $m = 12$ participants are plotted in Figure 1.[2]

Visual inspection and preliminary analysis suggested that it would be necessary to include time trends in any model for these data. In all of the following models, we use the full, piece-wise linear regression specification that allowed non-zero time trends; the models differ only in whether the within-case regression coefficients are assumed to be constant across individuals or are allowed to vary. In some cases, the effect size parameter depends on the choice of time-points A and B for describing the hypothetical between-subjects design. We use $A = 2$, meaning that in a hypothetical between-subjects experiment, the treatment would be introduced after the second measurement occasion. We also use $B = 9$, meaning that the effect size parameter measures the effect of $B - A = 7$ weeks of treatment, the maximum length observed in the data. To simplify the calculations, we center the weekly trend at $C = 9$ weeks, so that the within-case intercepts correspond to the average level of the outcome after 9 weeks, in the absence of treatment. We used the `nlme` package in `R` (Pinheiro et al., 2012) to obtain REML estimates.

We consider three different models for these data. An initial model assumes that the baseline time trends, the initial treatment effects, and the treatment-by-trend interaction are all constant across cases, but allows the baseline intercept (the average level of fatigue at week 9) to vary across cases. These assumptions correspond to MB 3. Table 1 reports REML estimates of the variance components, the fixed effects, and the effect size for this and following models. Also, Figure 1 plots predicted trends in the baseline and treatment phases for each case, generated using the best linear unbiased predictors of the random effects (for details, see

Pinheiro & Bates, 2000). Predicted trends based on MB3 estimates are plotted with solid lines. Based on these fitted trend lines and on formal comparisons to the other models under consideration, it is apparent that MB3 provides a poor fit; the assumption that baseline and treatment trends are constant across cases does not adequately describe these data.

Next, we consider allowing the baseline time trend to vary randomly across cases, which corresponds to MB4. REML estimates are reported in the second column of Table 1; predicted trend lines based on this model are plotted in Figure 1 using dotted lines. A likelihood ratio test comparing MB3 to MB4 rejects the simpler model (p = 0.002); visual inspection of the predicted trends also suggests an improved fit.

Based on the REML estimates from MB4, the intervention has a small immediate effect, lowering participants' fatigue scores by $\hat{\gamma}_{10} = -0.54$ scale points, followed by further decreases of $\hat{\gamma}_{30} = -1.63$ scale points per week. Combined, the treatment effect is -11.97 scale points after seven weeks of intervention. This treatment effect estimate is in the same units as the outcome measure; to convert it into an effect size, measured in standard deviation units, estimates of the model's variance components are needed. The REML estimate of within-case variance is $\hat{\sigma}^2 = 29$ squared scale points, assumed to be constant across measurement occasions. The between-case baseline variation is much larger: $\hat{\tau}_0^2 = 96$ squared scale points. Recall that the between-case variation is specific to a point in time, because MB4 assumes that the baseline trends vary across cases. Since we have centered time at week C = 9, $\hat{\tau}_0^2$ represents the between-case variation in the average level of the outcome at week 9, in the absence of treatment. Therefore, the total variation at week 9 is $\hat{\tau}_0^2 + \hat{\sigma}^2 = 125$, the square-root of which goes into the denominator of the effect size. We calculate the unadjusted effect size as

$\left(\hat{\gamma}_{10} + 7\hat{\gamma}_{30}\right) / \sqrt{\hat{\tau}_0^2 + \hat{\sigma}^2} = -11.97 / \sqrt{125} = -1.07$ standard deviation units. After using the estimated degrees of freedom to make a small-sample correction, the adjusted effect size is g_{AB} = -1.01 standard deviation units, with a standard error of 0.47. Using the symmetric *t* approach, a 95% CI for the effect size is [-2.02,-0.01]. Thus, despite the relatively large number of cases in the study, there remains considerable uncertainty about the magnitude of the design-comparable effect size.

We consider one further model for these data. MB5 assumes that baseline intercepts, baseline time trends, and treatment-by-time trend interactions all vary across cases, implying a covariance matrix for the random effects that has six parameters. With only twelve cases, use of REML estimation might be questionable if the goal were to draw inferences about the structure of the random effects. However, the simulation evidence that we have presented tentatively suggests that for purposes of obtaining a point estimate of the effect size, REML may be a reasonable strategy even with such a limited sample of cases.

The final column of Table 1 reports REML estimates for MB5. REML estimation of this model leads to an estimated correlation of very close to one between the random effects of trend and treatment-by-trend interaction (i.e., $\hat{\tau}_{32} / \sqrt{\hat{\tau}_2^2 \hat{\tau}_3^2} \approx 1$). The likelihood ratio test for MB5 versus MB4 is statistically significant (p = 0.030), though this asymptotic test may provide poor guidance with only twelve cases.

Though the estimated fixed effects are very similar to MB4, the estimated variance components and effect size are quite different. In particular, the between-case variation is only $\hat{\tau}_0^2 = 38$ squared scale points, compared to 96 for MB4, and the within-case variation is $\hat{\sigma}^2 = 23$ squared scale points, compared to 29 for MB4. The unadjusted effect size estimate is

$\hat{\delta}_{AB} = -1.50$ standard deviation units; after the degrees-of-freedom adjustment, g_{AB} = -1.33, with a standard error of 0.83 and a 95% symmetric *t* CI of [-3.29, 0.64]. The point estimate is much larger than the corresponding estimate from MB4, due mostly to the much smaller estimate of the between-case variation at week 9, which is only partially tempered by the reduced degrees of freedom. However, the precision of the estimate is extremely low. Given that this example uses a large number of cases for a single-case design, such substantial uncertainty should be anticipated when applying MB5 to single-case designs that display variable time trends.

Discussion

In studies evaluating the causal effects of interventions, an effect size provides a summary measure of the magnitude of an intervention's effect, in a metric that can be compared to effects from different studies or of other interventions. Ideally, an effect size should not be influenced by the incidental details of a study's design (such as sample size or pattern of treatment assignment), but should capture variability that is of scientific or policy interest. Comparability across multiple baseline designs and between-subjects experiments is thus an important criteria, as emphasized in several recent discussions of meta-analysis for single case research (Horner, Swaminathan, Sugai, & Smolkowski, 2012; Kratochwill et al., 2012).

This paper has provided a general framework for deriving design-comparable effect sizes for treatment effects in multiple baseline studies, which are on the same metric as Cohen's *d* from a between-subjects randomized experiment. The framework that we have described is both extensible, allowing effect size estimation under a variety of different models, as well as accessible, in that it can be implemented using standard software. The simulations studies provided initial evidence that the c-REML point estimator has reasonably small biases, even in

samples with very few cases. Larger numbers of cases are needed to obtain reasonably accurate standard errors and CIs when based on models that involve multiple random effects or extrapolation of time trends.

In meta-analytic contexts, inaccurate variance estimates will produce study-specific weights that are less than fully efficient and inaccurate estimates of between-study heterogeneity. Consequently, meta-analyses that use the proposed effect size and variance estimator should consider using robust standard errors (Hedges, Tipton, & Johnson, 2010) rather than relying on the model-based standard errors for the average effect size.

Given the idiographic orientation of single-case research, richly parameterized models with multiple components of variation are likely to be appealing on theoretical grounds, yet it is precisely these models that are difficult to fit with the number of cases encountered in many single-case studies. Other approaches to model fitting may offer some advantage and should be investigated further, including in particular fully Bayesian methods with weakly informative priors (Baldwin & Fellingham, 2013; Gelman, 2006) or related varieties of penalized likelihood methods (Chung, Rabe-Hesketh, Dorie, Gelman, & Liu, 2013). Parametric bootstrapping (Efron & Tibshirani, 1998, sec. 6.5) and small-sample corrections for the covariance matrix of fixed effect estimates (Kenward & Roger, 1997, 2009) warrant further investigation as well, as they may provide better means of estimating the effect size variance and constructing CIs.

However, more refined estimation procedures are no panacea. A fundamental limitation of the framework that we have presented is that design-comparable standardized mean differences are likely to be very imprecisely estimated when based on only small samples of cases. This is a consequence of the need to estimate a scale parameter (the denominator of the effect size) that includes between-case variance components. Another inherent consequence of

design comparability is that the operational definition of the target effect size depends on the choice of implementation time *A* and follow-up time *B*. Such arbitrary dependence may appear to be a drawback of the proposed effect size, but we would argue that it may instead indicate that the broader modeling assumptions (such as the inclusion of treatment-by-time trend interactions) need to be re-evaluated relative to the context of the analysis.

The simulation studies that we have presented are limited in several ways, the most obvious being that we have examined only a fairly small set of models and, for models with more than a single variance component, only a subset of the full parameter space. Furthermore, we considered only one particular method for evaluating the covariance matrix of the variance component parameters, using an explicit formula for the expected information matrix of the restricted likelihood. The exact bias of the adjusted REML effect size estimator depends on this covariance (through the degrees of freedom correction), yet software packages differ in how the covariance matrix is estimated. When presenting the results of REML model estimation and effect size calculations, it would therefore prudent to also report the software version used to fit the model and the method used to estimate the covariance matrix of the variance component parameters.

While this paper has focused on design-comparable effect sizes, we acknowledge that design-comparability is not the only or over-riding desideratum of effect sizes for use in single case research. Other effect sizes, including ones that are not identified in between-subjects designs, may be as useful or more useful as summaries of treatment effects from single-case studies. If scientific theory or empirical evidence suggests that a different metric better quantifies the underlying phenomenon being studied, its lack of design-comparability should not prevent its consideration. Still, we expect that the models and methods discussed in this paper will be useful

analytic tools for defining and estimating many sorts of effect sizes, and especially for understanding the comparability of effect sizes across different research designs.

Notes:

1. In the notation of Hedges et al. (Hedges et al., 2013), $\gamma_{00} = \mu^C$, $\gamma_{10} = \mu^T - \mu^C$, $\eta_{0i} = \eta_i$, and $\tau_0^2 = \tau^2$.
2. The raw data are reported in Table 1 of Schutte et al. (Schutte et al., 2008). R code for the calculations reported in this section is available in the package `scdhlm`, which is included in the supplementary materials.

References

Algina, J., & Keselman, H. J. (2003). Approximate confidence intervals for effect sizes. *Educational and Psychological Measurement*, *63*(4), 537–553. doi:10.1177/0013164403256358

Bailey, J. S., & Burch, M. R. (2002). *Research Methods in Applied Behavior Analysis*. Thousand Oaks, CA: Sage Publications.

Baldwin, S. A., & Fellingham, G. W. (2013). Bayesian methods for the analysis of small sample multilevel data with a complex variance structure. *Psychological Methods*, *18*(2), 151–64. doi:10.1037/a0030642

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

Borckardt, J. J., Nash, M. R., Murphy, M. D., Moore, M., Shaw, D., & O'Neil, P. (2008). Clinical practice as natural laboratory for psychotherapy research: a guide to case-based time-series analysis. *The American Psychologist*, *63*(2), 77–95. doi:10.1037/0003-066X.63.2.77

Busse, R. T., Kratochwill, T. R., & Elliott, S. N. (1995). Meta-analysis for single-case consultation outcomes: Applications to research and practice. *Journal of School Psychology*, *33*(4), 269–285.

Center, B. A., Skiba, R. J., & Casey, A. (1985). A methodology for the quantitative synthesis of intra-subject design research. *The Journal of Special Education*, *19*(4), 387–400.

Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A., & Liu, J. (2013). A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika*, *78*(4), 685–709. doi:10.1007/s11336-013-9328-2

Cooper, H. M. (2009). Hypotheses and problems in research synthesis. In H. M. Cooper, L. V Hedges, & J. Valentine (Eds.), *The Handbook of Research Synthesis and Meta-Analysis* (2nd ed., pp. 19–35). New York, NY: Russell Sage Foundation.

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, *61*(4), 532–574. doi:10.1177/0013164401614002

Efron, B., & Tibshirani, R. J. (1998). *An Introduction to the Bootstrap*. Boca Raton, FL: Chapman & Hall/CRC.

Ferron, J. M., Bell, B. A., Hess, M. R., Rendina-Gobioff, G., & Hibbard, S. T. (2009). Making treatment effect inferences from multiple-baseline data: The utility of multilevel modeling approaches. *Behavior Research Methods, 41*(2), 372–84. doi:10.3758/BRM.41.2.372

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis, 1*(3), 515–533.

Gorsuch, R. L. (1983). Three methods for analyzing limited time-series (N of 1) data. *Behavioral Assessment, 5*(2), 141–154.

Gottman, J. M. (1981). *Time-Series Analysis: A Comprehensive Introduction for Social Scientists*. Cambridge, England: Cambridge University Press.

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

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

Hedges, L. V. (2011). Effect sizes in three-level cluster-randomized experiments. *Journal of Educational and Behavioral Statistics, 36*(3), 346–380. doi:10.3102/1076998610376617

Hedges, L. V, Pustejovsky, J. E., & Shadish, W. R. (2012). A standardized mean difference effect size for single case designs. *Research Synthesis Methods*, *3*, 224–239. doi: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. doi:10.1002/jrsm.1086

Hedges, L. V, Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1*(1), 39–65. doi:10.1002/jrsm.5

Hersen, M. (1990). Single-case experimental designs. In A. S. Bellack, M. Hersen, & A. E. Kazdin (Eds.), *International Handbook of Behavior Modification and Therapy* (2nd ed., pp. 175–210). New York, NY: Plenum Press.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960.

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

Horner, R. H., Swaminathan, H., Sugai, G., & Smolkowski, K. (2012). Considerations for the systematic analysis and use of single-case research. *Education and Treatment of Children*, *35*(2), 269–290. doi:10.1353/etc.2012.0011

Huitema, B. E., & McKean, J. W. (2000). Design specification issues in time-series intervention models. *Educational and Psychological Measurement*, *60*(1), 38–58. doi:10.1177/00131640021970358

Kazdin, A. E. (2011). *Single-Case Research Designs: Methods for Clinical and Applied Settings*. New York, NY: Oxford University Press.

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, *53*(3), 983–97.

Kenward, M. G., & Roger, J. H. (2009). An improved approximation to the precision of fixed effects from restricted maximum likelihood. *Computational Statistics & Data Analysis*, *53*(7), 2583–2595. doi:10.1016/j.csda.2008.12.013

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2012). Single-case intervention research design standards. *Remedial and Special Education*, *34*(1), 26–38. doi:10.1177/0741932512452794

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. New York, NY: Springer Verlag.

Pinheiro, J. C., Bates, D. M., DebRoy, S., & Sarkar, D. (2012). nlme: Linear and Nonlinear Mixed Effects Models. Retrieved from http://cran.r-project.org/package=nlme

SAS Institute Inc. (2008). *SAS/STAT(R) 9.2 User's Guide*. Cary, NC: SAS Institute Inc.

Schutte, N. S., Malouff, J. M., & Brown, R. F. (2008). Efficacy of an emotion-focused treatment for prolonged fatigue. *Behavior Modification*, *32*(5), 699–713. doi:10.1177/0145445508317133

Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The quantitative synthesis of single-subject research. *Remedial and Special Education*, *8*(2), 24–43.

Shadish, W. R., Rindskopf, D. M., Hedges, L. V, & Sullivan, K. J. (2013). Bayesian estimates of autocorrelations in single-case designs. *Behavior Research Methods*, *45*(3), 813–821. doi:10.3758/s13428-012-0282-1

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

Singer, J. D., & Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. New York, NY: Oxford University Press.

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

StataCorp. (2011). *Stata, Release 12: Longitudinal-Data/Panel-Data Reference Manual*. Stata Press.

Wolery, M., Busick, M., Reichow, B., & Barton, E. E. (2010). Comparison of overlap methods for quantitatively synthesizing single-subject data. *The Journal of Special Education, 44*(1), 18–28. doi:10.1177/0022466908328009

Table 1
Model estimates for Schutte, et al. (2008) data

Parameter	MB3 Estimate	(s.e.)	MB4 Estimate	(s.e.)	MB5 Estimate	(s.e.)
Variance components						
Within-case var. ($\hat{\sigma}^2$)	99.00	(47.42)	29.39	(6.67)	22.54	(4.38)
Autocorrelation ($\hat{\phi}$)	0.81	(0.09)	0.40	(0.13)	0.24	(0.14)
Between-case var. ($\hat{\tau}_0^2$)	14.77	(44.39)	95.71	(46.94)	38.30	(33.49)
Case-trend covariance ($\hat{\tau}_{20}$)			11.21	(6.37)	0.38	(3.51)
Trend variance ($\hat{\tau}_2^2$)			1.99	(1.08)	0.15	(0.57)
Case-Trt. × Trend cov. ($\hat{\tau}_{30}$)					1.74	(7.59)
Trend-Trt. × Trend cov. ($\hat{\tau}_{32}$)					0.67	(0.99)
Trt. × Trend variance ($\hat{\tau}_3^2$)					3.01	(2.74)
Total variance ($\hat{\tau}_0^2 + \hat{\sigma}^2$)	113.77	(29.80)	125.11	(46.78)	60.84	(32.88)
Fixed effects						
Intercept ($\hat{\gamma}_{00}$)	52.93	(4.42)	50.29	(4.07)	50.53	(2.82)
Treatment ($\hat{\gamma}_{10}$)	-1.37	(1.97)	-0.54	(1.75)	0.03	(1.60)
Weekly trend ($\hat{\gamma}_{20}$)	0.49	(0.62)	0.20	(0.62)	0.22	(0.36)
Trt. × Trend ($\hat{\gamma}_{30}$)	-1.90	(0.94)	-1.63	(0.66)	-1.67	(0.74)
Trt. effect after 7 weeks ($\mathbf{p}^T\hat{\boldsymbol{\gamma}}$)	-14.65	(6.34)	-11.97	(4.61)	-11.67	(5.18)
Effect size						
Unadjusted ($\hat{\delta}_{AB}$)	-1.37	(0.64)	-1.07	(0.50)	-1.50	(0.93)
Adjusted (g_{AB})	-1.34	(0.63)	-1.01	(0.47)	-1.33	(0.83)
Degrees of freedom (ν)	29.15		14.30		6.85	
Constant κ	0.59		0.41		0.66	
Log-likelihood	-435.09		-428.98		-424.51	

Figure 1
Multiple baseline data from Schutte, et al. (2008)