# Effect Size Estimation and Synthesis of Single-Case Designs: A Methods Guide

John M. Ferron, Lodi Lippen, Megan Kirby, Wendy Machalicek, James Pustejovsky, and Man Ch

2022-02-28

# Contents

# Preamble

Title page

Prepared for IES acknowledgement

Prepared by authors

Contract no

Advisory board

# Chapter 1

# Approaches for Effect Size Estimation and Synthesis of Single-Case Designs

This chapter provides the background and purpose for this methods guide. It also gives an overview of the three major approaches for estimation and synthesis of single-case studies:

a) design-comparable effect sizes,
b) case-specific effect sizes, and
c) multilevel modeling of the raw individual participant interrupted time-series data.

We describe the motivation and rationale for each of these approaches and provide a series of decision rules to guide researchers in selecting among them. The remaining chapters provide details and illustrations of the methods.

## 1.1 Background

Educational decisions at the state, district, school, and student level are now expected to be informed by empirical evidence (**??**). These expectations create a major need for synthesis, or the integration of research findings from multiple, existing sources of evidence—including findings from single-case designs (SCDs). (**??**)y developments in education related to the documentation of evidence-based practices. The first is the expanded use of SCD research methods across varying disciplines over the last 50 years, in both general and special education contexts

(**?**). Although the history of SCD methodology is beyond our present scope, innovations in design have expanded the use of SCDs across a range of professional fields, moving beyond SCD's quasi-experimental and behavior analysis origins (see **?**, for more detailed information on the history of SCD). Over the past 20 years, researchers' commitment to using rigorous procedures to identify evidence-based educational practices affirms not only the importance of randomized control trials (RCTs), but the effectiveness and efficiency of SCD.

The second development is the emergence of effect size (ES) measures and synthesis methods for use with SCD and other interrupted time series data (**??**). Interest in synthesis of SCDs is long-standing (e.g., **???**), emerging around the same time as methods for meta-analysis of group designs were becoming more statistically rigorous and sophisticated (**??**). However, the past twenty years have seen increased intensity of methodological research focused on SCDs and a substantial expansion in the diversity, flexibility, and accessibility of available analytic methods upon which researchers can draw. We view the current state of methodology as falling into three strands:

a) approaches that summarize the effect of each case and then synthesize these case-specific effect sizes (e.g., **?**),
b) methods that use multi-level models for analyzing the raw (or standardized) data from one or multiple SCD studies (**?**), and
c) techniques that use design-comparable effect size metrics (**?????**).

The present guide is organized around these three broad methodological strands.

The third development is the increased use of systematic review and meta-analysis procedures to identify and affirm evidence-based practices in education (**??**). Maggin and colleagues (**?**) examined the production of systematic reviews and meta-analyses of SCDs from 1985 to 2009, finding a marked increase in the appearance of such reviews between 2000 and 2009. In another survey, Jamshidi and colleagues (**?**) found increasing production through 2015, as well as improvements in the quality of published reviews. However, they also noted that, even in more recent literature, reviews often frequently failed to use appropriate methods for combining findings across studies (**?**). Thus, there remains a need for guidance about how to select and apply methods for synthesizing findings from SCD research.

Effect size (ES) measures are a critical companion to visual analysis for the interpretation of single-case research results, and are a key input in two of the available approaches for meta-analytic synthesis of SCDs. Although a variety of technically sound ES metrics exist for researchers to use when interpreting SCD findings, relatively few published meta-analyses use these more advanced techniques (cf. **???**). One reason for the lack of widespread use by researchers may be the conceptual and procedural complexity associated with advances in ES measures and meta-analysis techniques. A more rapid uptake of ES estimation methods may be hindered by the complexity of data extraction and calculation of ES for SCD. In addition, researchers may lack software tools for ES estimation for single-case studies when using some common statistical packages such as SPSS and SAS (**???**).

## 1.2 Purpose of the Methods Guide

The purpose of this methods guide is to improve educational researchers' practice in the estimation of ES measures and synthesis of findings from SCDs. We recognize that no single method is ideal for all research goals. Furthermore, methods that have the most to offer can be complex and may appear difficult to carry out. Thus, through the use of this methods guide, we aim to: (a) provide guidance and decision rules to simplify the process of selecting effect size estimation and synthesis methods and (b) illustrate the step-by-step application of appropriate methods using readily available tools, such as web-based software to calculate case-specific ES or design-comparable ES for SCD studies (e.g., **?**).

## 1.3 Study Quality

Conducting a research synthesis (whether of group design studies, SCDs, or both) involves several stages, starting with formulating the research aims, specifying inclusion criteria, conducting a systematic search, and screening for eligible studies (**??**). Additionally, before carrying out effect size calculations or meta-analysis, it is critical to consider the methodological quality and potential biases of studies to be included in the synthesis. To assist researchers in doing so, several distinct sets of standards are available for SCD studies (e.g., **????**). After assessing methodological quality, researchers can use one of two strategies to guide the estimation and synthesis of effect sizes. One strategy incorporates consideration of study quality as an inclusion criteria, so that low-quality studies are screened out and synthesis is based on the subset of studies with quality high enough so that changes in outcome(s) can reasonably be attributed to the intervention. For example, the What Works Clearinghouse Single-Case Pilot Standards (**?**) indicated that effect sizes are only to be computed after studies have been shown to meet both minimum design criteria (e.g., a multiple-baseline study has at least three temporally spaced opportunities for the effect to be demonstrated, along with phases of at least three observations) and minimum evidence criteria (e.g., visual analysis of the data from the study show experimental control so that the changes in the outcome can be attributed to the intervention). Concerns with estimating effect sizes for SCD studies without the use visual analysis to rule out alternative explanations for observed changes in the outcomes continues to be echoed in the literature (**??**). An alternative strategy in considering study quality is to use broader inclusion criteria, code for aspects study quality, and examine study quality codes as potential moderators in a meta-analysis. With this approach, researchers can estimate ESs for studies of varying degrees of quality and empirically explore whether the magnitude of ESs varies depending on aspects of study quality.

We assume that researchers who use this methods guide will have already selected a method to assess study quality and an approach for incorporating those assessments into their synthesis. As such, we do not focus on SCD study quality assessment methods, but rather provide guidelines that can be applied to a collection of studies that have met the researchers' inclusion criteria, which potentially include criteria related to study quality. Thus, in this guide we focus on the final stages of a research synthesis, on the questions of how to select a method for estimating effects, how to compute ES estimates (or otherwise prepare data for synthesis), and how to synthesize findings in the form of ES estimates or individual-level data.

## 1.4 Selecting an Approach for Effect Estimation and Synthesis

In order to select an approach for estimating and synthesizing effects from SCDs, we recommend that researchers first reflect on the research aims that motivate their synthesis. In some contexts, researchers' primary aims may be focused on summarizing evidence to arrive at statements about average efficacy of a class of interventions. In other contexts, researchers might instead or additionally be interested in understanding variation in effects and the extent to which such variation is associated with characteristics of participants, specific features of interventions, or other contextual factors. When the focus is mostly on summarization, researchers may find it more useful to use design-comparable effect sizes that describe average effects. If individual-level variation is the focus, then approaches using case-specific effect sizes or multi-level modeling may be more advantageous.

Another over-arching consideration for selecting a synthesis approach pertains to the features of the studies identified for inclusion in the synthesis. Quantitative synthesis requires choosing an effect size metric that permits comparisons of the magnitude of effects across individual participants and studies. Consequently, the extent to which eligible studies use different types of designs or different types of outcome measurements creates constraints on how effects from those studies can be described or compared. For instance, if all eligible studies used multiple baseline designs across participants (or another common type of SCD), then several different synthesis approaches are feasible. In contrast, if eligible studies include both SCDs and group design studies (such as small randomized experiments, each with a single pre-test and a single post-test), then researchers need a synthesis approach that permits comparisons across both types of designs. Similarly, if all eligible studies used SCDs with very similar methods for assessing the dependent variable, then synthesis based on multi-level modeling of raw data is possible. In contrast, if eligible studies used a variety of different assessments, then a synthesis approach based on case-specific effect sizes may be required.

These two broad considerations—the aims of the synthesis and the features of eligible studies—can guide the selection of an approach for synthesis of SCDs. We now describe in more detail how three broad approaches to synthesis fit within these considerations.

### 1.4.1 Design-Comparable Effect Sizes

In some situations, the aim of the research team is to synthesize the evidence for intervention effectiveness from both single-case and group design studies. For example, a meta-analysis by Wood and colleagues (**?**) analyzed 22 single-case and between-group studies to examine the effects of text-to-speech and other read-aloud tools on reading comprehension outcomes for students with reading disabilities. The authors used the standardized mean difference to estimate read-aloud intervention effects in the group design studies and a comparable standardized mean difference from the included single-case research, resulting in an overall average weighted effect size of $d = 0.35$, 95% CI (0.14, 0.56). Because the purpose of the **?** study involved the comparison and averaging of effects across single-case and group designs, it was critical to use an ES metric that is theoretically comparable across the designs. In similar situations, researchers should select from the design-comparable effect size options (**??????**). However, if researchers aim to synthesize findings from only single-case studies (i.e., not to integrate findings across single-case and group design studies), other options may be preferable.

### 1.4.2 Case-Specific Effect Sizes

In addition to summarizing average effects across studies, researchers may also be interested in exploring variation in treatment effects across individual participants. When dependent variables are measured differently across studies, it is important for researchers to use an effect size metric and synthesis approach that accounts for such. For example, Bowman-Perrott and colleagues (**?**) examined five potential moderators of the effectiveness of the Good Behavior Game in promoting positive behavior in the classroom. Results of their meta-analysis suggested that the intervention was most effective in reducing problem behaviors among students with or at risk for emotional and behavioral disorders. Another meta-analysis by Mason and colleagues (**?**) investigated the moderating effects of participant characteristics, targeted outcomes, and implementation components on the efficacy of video self-modeling, in which a learner with disabilities watches a video of a model engaged in targeted skills or tasks. They found that intervention effects were stronger for younger participants with autism-spectrum disorders.

Because these syntheses focused on investigating variation across individuals in the effect of treatment, it was important that the effect size estimation and synthesis approach produced effect estimates for each individual participant, rather than a study-level summary effect estimates. In addition, because the outcome measure differed among studies, the researchers needed an individual-level effect size metric that is not scale-dependent (e.g., they could not be based on simple raw score mean differences). In contexts like these, researchers should consider selecting among the case-specific effect size estimation and synthesis options. The case-specific effect size options are not viable for synthesis across single-case and group design studies because the case-specific effects are not comparable to the group design effects. However, the design-comparable effects are not viable for studying within-participant variation between individuals in treatment effectiveness because they produce effect estimates at the study level (i.e., the effect averaged across the cases), not the individual level.

### 1.4.3 Multilevel Modeling of Individual Participant Interrupted Time-Series Data

Finally, researchers might have identified a set of SCDs that all use the same or very similar outcome measures, with the aim of studying the variation in effects over time within and between individuals. For example, Datchuk and colleagues (**?**) meta-analyzed 15 single-case studies with 79 students to examine the effects of an intervention on the level and trend in correct writing sequences per minute for students with disabilities. They found the effect increased with time in intervention (i.e., there was a positive effect on the trend) and that this temporal change in effect was more pronounced with younger students. When focusing on both variation in effect over time and variation in effect across cases, it is important that the researchers select a meta-analytic approach that does not rely only on a single effect estimate for a study (e.g., design-comparable effect sizes) or even a single effect estimate for a case (e.g., case-specific effect sizes). In contexts like this, we suggest researchers consider options for multilevel modeling of individual participant data series.

### 1.4.4 Summary of Options for Effect Estimation and Synthesis

The flow chart in Figure **??** illustrates a set of heuristic decision rules for selecting among the three general approaches to estimating and synthesizing single-case research. If the primary purpose of one's research is to integrate findings from both single-case and group-design studies, the researcher should consider design-comparable effect sizes. Alternately, if the primary purpose of the research is to integrate findings from SCDs only, additional questions should be addressed related to the measurement of the dependent variable. If the outcome of interest is measured in different ways for different cases and the researcher aims to examine how effects vary across the cases, then the researcher should consider the options for estimating and synthesizing case-specific effect sizes. If the outcome is measured the same way across cases, and there is interest in how the effect changes over the course of an intervention, the researcher should consider multilevel modeling of the raw data series.

### 1.4.5 Limitations in Selecting an Approach for Effect Estimation and Synthesis

We emphasize that Figure **??** presents a heuristic, simplified procedure for the selection among the three general approaches to ES estimation and synthesis, which does cover every possible research context. There will surely be situations where researchers' aims and contexts differ from those we describe, and thus do not align perfectly with one of our primary approaches to estimating and synthesizing single-case effect sizes. For example, researchers who are synthesizing findings from a set of SCDs may wish to compare their results to a previously published meta-analysis of group design studies, but not to investigate individual-level variation in treatment effects. They may therefore elect to use design-comparable effect sizes even though they are not formally integrating results from group design studies within their review.

A further possibility is that researchers might elect to use multiple approaches to synthesis in order to address different aims or questions. For example, consider a project in which researchers have identified both single-case and group design studies. They might want to integrate findings across design types while also exploring the variation in effects among individuals. In this scenario, researchers could estimate design-comparable effect sizes for their first aim and case-specific effect sizes from the subset of single-case studies for their second aim.

We also note that there may be situations that falls in between those we described for case-specific effect sizes and those for multilevel modeling of the raw data series. For example, researchers may want to examine how effects vary over time and across cases, using studies with different outcomes. For this purpose, the researchers can use extensions of the primary approaches we present. The researchers could either standardize the raw data before estimating a multilevel model, or they could synthesize case specific effect sizes where they use multiple standardized effects for each case (e.g., an effect that indexes the immediate shift in level, and another effect that indexes a change in slope). Our simplified selection method cannot exhaustively cover all currently available options.

Finally, we anticipate that the heuristic guidance we provide here will need to be refined over time, as further methodological innovations become available. We anticipate that research presently underway will provide even more meta-analytic options in the future, with implications for how to select an approach for synthesis. At some point it may be possible to compute case-specific effect sizes that are also design-comparable, or it may be possible to standardize the

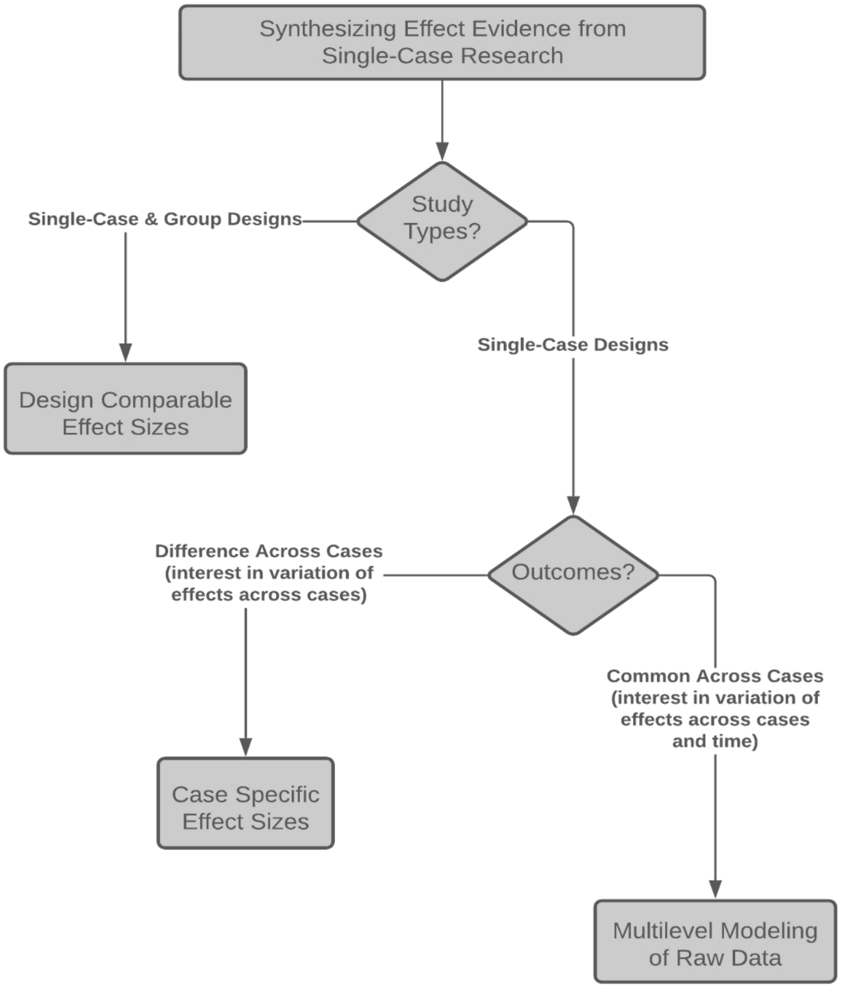
Synthesizing Effect Evidence from Single-Case Research
Study Types?
Single-Case & Group Designs
Design Comparable Effect Sizes
Single-Case Designs
Outcomes?
Difference Across Cases (interest in variation of effects across cases)
Case Specific Effect Sizes
Common Across Cases (interest in variation of effects across cases and time)
Multilevel Modeling of Raw Data

Figure 1.1: Flow chart showing the approach for synthesizing effect evidence based on the types of studies being examined and whether the outcome variable is common across cases.

data for multilevel models in a way that leads to parameter estimates from the model that correspond to design-comparable effect estimates. If such methods become available, some of the distinctions made here will become artificial. In the time, remainder of this methods guide follows the proposed heuristics for selecting among the three major approaches to effect size estimation and synthesis. Even as methodology continues to advance, researchers need guidance that acknowledges the complexity of research purposes and contexts and is dynamic in its accommodation of such variation, while also being concrete and straightforward enough to be put into practice.

## 1.5 Structure of the Methods Guide

We divide the remainder of this guide into three major sections:

a) design-comparable effect size estimation and synthesis,
b) case-specific effect size estimation and synthesis, and
c) multilevel modeling to estimate and synthesize effects.

These sections do not build upon each other or need to be read in order. Rather, we expect those using this guide to follow the decision rules in Figure **??** to determine which section of the guide is most appropriate for them, and then to jump immediately to that section.

Each major section is divided into chapters. The initial chapter of each section introduces the specific approach and its assumptions, discusses when to use it, and what options exist within the approach. Furthermore, we provide additional decision rules for selecting among the specific techniques and options available with a given broad approach. We encourage researchers to use the decision rules within the initial chapter of the major section to get to an appropriate option. Then, researchers can directly refer to the chapter that has the illustration for that specific option. For each illustration, we:

1) describe the purposes for estimating and synthesizing effects, and the available data,
2) demonstrate how to use the decision rules in Figure **??**, along with the additional decision rules within the initial chapter of the major section, to arrive at the option being illustrated,
3) present the data for the illustration showing how it needs to be structured for the analysis, and
4) provide a step-by-step illustration of how to estimate and synthesize effects using readily available analysis tools.

# Chapter 2

# Introduction to Design-Comparable Effect Sizes

This chapter provides background on design-comparable effect sizes, describes when to use them, and explains the key assumptions behind their use. In particular, we highlight both the assumptions for estimation of effect sizes and assumptions for using the effect sizes in meta-analytic synthesis. We then describe options that are available for estimating design-comparable effect sizes. These options allow for different assumptions regarding trends in baseline or treatment phases, as well as different assumptions about the variation in the treatment effect across cases. We close by providing a set of decision rules for selecting among the options for design-comparable effect sizes.

## 2.1 Background

Design-comparable effect sizes are effect size (ES) indices for single case studies that are on the same metric as ES indices used in group comparison studies (**??????**). These design-comparable ESs are valuable in a variety of synthesis contexts, particularly those that involve both single case designs (SCDs) and group comparison studies. However, these methods also have several limitations. First, because design-comparable ESs rely on an estimate of the variance between participants, they require at least three distinct participants for estimation, which limits their use to SCDs involving multiple participants (such as multiple baselines across participants) or close replications of a design, each of which has a single participant (such as an ABAB design replicated across several students). The methods are also only available for multiple baseline, multiple probe, or treatment reversal designs. Extensions for alternating treatment designs and other types of SCDs have yet to be developed.

A second, conceptual limitation, is that design-comparable ES produce a single, summary ES per outcome, which represents an average effect across participants—just as ES from between-group designs are summaries of average effects. As a result, the design-comparable ES might conceal heterogeneity of effects across the participants in an SCD. Thus, when using such ES in meta-analysis, moderator analyses are limited to study-level characteristics (i.e., no examination of potential moderators that vary across cases in a study or across time points within a case).

A third limitation is that design-comparable ES are based on a hierarchical model that involves specific assumptions about the distribution of outcomes measured in the study. Developing a reasonable model requires care and attention to the plausibility of its assumptions, and is not an automatic process. Moreover, for some types of outcomes, the distributional assumptions of the model may not be appropriate, which creates a further limitation on use of design-comparable ES.

In this chapter, we describe when researchers should select to use design-comparable ES estimates, give a precise definition of the design-comparable standardized mean difference ES, and explain the assumptions made when using this approach for effect estimation and synthesis. We then provide six of the most common modeling options and guidance to assist researchers with the most appropriate selection of an option.

## 2.2 When to Use Design-Comparable Effect Sizes

When choosing an effect size (ES) for single case design (SCD) data, the broader purpose for computing ESs needs to be considered first and foremost. In some cases, researchers may want to synthesize results across different types of studies. For example, they may want to estimate the size of the effect based on the data available from all studies in an area (e.g., all studies examining the effects of social skills intervention on the social and academic outcomes of elementary aged students with disabilities). In some areas of educational research, the literature identified for synthesis may include both group and single case experimental studies. To average the effect across studies with different designs, one must pick an ES index that has a comparable interpretation for each of the included designs (**?????**). Researchers developed the design-comparable ES for SCDs for exactly this purpose. It provides an ES on a common metric by answering the question "What would the standardized mean difference ES be if one could have somehow performed a between-group randomized experiment based on the same population of participants, same intervention protocol, and same outcome measures?" Design-comparable ESs can be computed for multiple-baseline designs across 3 or more participants, multiple-probe designs across 3 or more participants, and reversal (e.g. ABAB) designs replicated across 3 or more participants. Ongoing research is likely to increase the designs for which design-comparable ESs can be estimated.

## 2.3 A General Definition of Design-Comparable Effect Sizes

To understand the logic of the design comparable ES, it is helpful to first consider how effect sizes are defined in group design studies. In a between-groups randomized experiment comparing an intervention condition (B) to a control condition (A) for a specified population, results are commonly summarized with the standardized mean difference (SMD) ES index. The SMD effect size parameter can be defined as

$$\delta = \frac{\mu_T - \mu_C}{\sigma_C}, \tag{2.1}$$

where $\mu_T$ is the average outcome if the entire population were to receive the intervention, $\mu_C$ is the average outcome if the entire population were to receive the control condition, and $\sigma_C$ is the standard deviation of the outcome if the entire population were to receive the control condition. The effect size may be estimated by substituting sample means and sample standard deviations in place of the corresponding population quantities (**?**), or by pooling sample standard deviations across the intervention and control conditions under the assumption that the population variance is equal. Alternately, the mean difference in the numerator of the effect size may be estimated based on a statistical model, such as an analysis of covariance that adjusts for between-group differences on baseline characteristics (**?**). Researchers often apply the Hedges $g$ small-sample correction, which reduces the bias of the effect size estimator that arises from estimating $\sigma_C$ based on a limited number of observations (**?**).

The design-comparable SMD for SCDs aims to estimate the same quantity as the SMD from a between-groups experiment given in Equation (**??**), using data from a multiple baseline, multiple probe, or replicated treatment reversal design. Doing so is made more challenging, however, because the data from such SCDs involves repeated measurements taken over time. In order to precisely define a design-comparable SMD, we must therefore be specific about the timing of intervention and outcome assessment. Hypothetically, if we could conduct a between-groups experiment using the same study procedures as the SCD, we would still need to decide when to begin intervention and when to collect outcome data. Suppose that the SCD takes place over times $t = 1, ..., T$. In our hypothetical between-group experiment, we start intervention at time $A$, for $1 \leq A < T$ and we collect outcome data for all participants at time $B$, for

$A < B$. The SMD from such an experiment would then correspond to

$$\delta_{AB} = \frac{\mu_B(A) - \mu_B(T)}{\sigma_B(T)}, \tag{2.2}$$

where $\mu_B(A)$ is the average outcome at follow-up time $B$ if the entire population were to receive the intervention at time $A$, $\mu_B(T)$ is the average outcome at follow-up time $B$ if the entire population were to receive the intervention at time $T$, and $\sigma_B(T)$ is the standard deviation of the outcome at follow-up time $B$ if the entire population were to receive the intervention at time $T$. Note that $\mu_B(T)$ corresponds to the average outcome under the control condition ($\mu_C$ above), because participants would not yet have received intervention as of time $B$. Similarly $\sigma_B(T)$ is the analogue of $\sigma_C$, the standard deviation of the outcome under the control condition, because participants would not yet have received the intervention as of time $B$.

**?** described a strategy for estimating the design-comparable SMD ES $\delta_A B$, using data from an SCD study. Broadly, the strategy involves specifying a multi-level model for the data, estimating the component quantities $\mu_B(A)$, $\mu_B(T)$, and $\sigma_B(T)$ based on the specified model, and applying a small-sample correction analogous to Hedges' $g$. This strategy will only work if the SCD study includes data from multiple participants because we need to be able to estimate $\sigma_B(T)$, the standard deviation of the outcome across the population of participants. The strategy involves a multi-level model for the data because SCDs involve repeated measurements on a set of participants. Thus, the first level of the model describes the pattern of repeated measurements over time, nested within a given participant, and the second level of the model describes how the first-level parameters vary across participants. As a result, the model involves decomposing the standard deviation of the outcome $\sigma_B(T)$ into within-participant variation and between-participant variation. This decomposition is not typically possible in a between-groups randomized experiment (unless researchers collect repeated measures of the outcome for each participant).

## 2.4 What We Assume with Design-Comparable Effect Size Estimation and Synthesis

Design-comparable ESs require assumptions for estimation from SCD single case data, as well as further assumptions for inclusion in a synthesis. Regarding synthesis, we assume the effect sizes from all studies are exchangeable, meaning that they are similar (though not necessarily identical) and that their ranking cannot be systematically predicted. Regarding estimation, design-comparable ESs for SCD studies rely on multilevel models with normally distributed error terms. Thus, there are assumptions about the underlying structural model (e.g., whether or not there are trends in phases), as well as assumptions about the error terms (e.g., errors are normally distributed and homoscedastic). We now explain these assumptions in greater detail.

### 2.4.1 Exchangeable Effect Sizes

The predominant statistical model for synthesizing effect sizes is known as the random effects model, the assumptions of which can be motivated in several different ways. One way is to imagine that the studies included in a synthesis comprise a random sample from a super-population of possible studies on the topic of interest. The model can also be motivated in a Bayesian framework by the assumption of exchangeability, meaning that the effect sizes of studies included in a synthesis are on a common metric that permits judgements of similarity and that their relative magnitude cannot be systematically predicted a priori (**?**). For brevity, we refer to both the super-population and Bayesian motivations as the exchangeability assumption. Crucially, the exchangeability assumption depends on the effect size metric used for synthesis. In other words, for a given set of studies, the assumption may be reasonable for one effect size metric but not reasonable for another.

Design-comparable ESs for SCDs have thus far been defined for the standardized mean difference metric, which describes the magnitude of an intervention effect in terms of a mean difference, standardized by the population standard deviation of the outcome in the absence of treatment. For the SMD metric, exchangeability is more plausible when the population of participants from one study is similar—or even interchangeable with—the population of participants from another study. If the populations are similar, and if two studies used the exact same operational measure of the dependent variable, then the distribution of outcomes in the control condition should have similar variance. In contrast, suppose that the two studies draw from populations with very different characteristics, so that the distribution of the dependent variable in one study population is much less dispersed than the distribution of the same variable in the other study population. In this scenario, an intervention that produces identical effects on the scale of the dependent variable would nonetheless have quite different effect sizes on the scale of the SMD, making the exchangeability assumption less tenable.

This aspect of the exchangeability assumption can be explored by examining the sampling methods and measurement procedures used in the studies to be synthesized. In particular, when subsets of studies use the same operational measure of the dependent variable, the between-participant variance in those studies can be compared. To illustrate the exploration of the variability between cases, consider the sampling procedures used in the following two studies which examine the effect of interventions on writing performance as measured by correct word sequences. In one study, the sample consisted of three 7-year-old white males identified by their teachers as struggling with writing (**?**). In the other study (**?**), the sample consisted of three 10-year-old students who exhibited poor writing skills. The first student was an African American male identified with an emotional disturbance, the second student was a white male identified with a specific learning disability, and the third student was a while female identified with a specific learning disability. Presented with this information, we ask ourselves the following questions: Are these samples similar enough to satisfy the exchangeability assumption with the SMD metric? Or might the second sample, which included older students and variability across race, gender, and disability, be so much more variable in writing performance that it is not reasonable to use the SMD metric to judge similarity of effects?

In the first study (**?**), the mean number of correct word sequences during baseline for the three participants were 8.29, 15.0, and 10.8. In the second study (**?**), the mean baseline levels were 14.6, 29.1, and 22.1. In the first study, the variation between cases, as indexed by the standard deviation (SD) of the three baseline means, is 3.4, whereas in the second study this between case SD is 7.3. Is this difference large enough to distort the design-comparable ES?

To address our questions, we first consider the raw score ES for each study. In Study 1, the shift in the expected number of correct word sequences when moving from baseline to intervention is 10.7 (based on a multilevel model that assumes no trends in baseline or treatment phase, and variation in the effect across cases). However, in Study 2 the raw score ES is 9.6 correct word sequences. Thus, the raw score ES from the first study is about 1.1 times larger than the ES we find in the second study. Next, we consider the design-comparable ES computed based on a model that assumes no trends in either baseline or treatment phases and demonstrates variability in the treatment effect across cases just like the model for the raw score ES. For the first study, the design-comparable ES is 0.962, and in the second study it is 0.827; thus the first study's ES is about 1.2 times the second's, similar to what was seen with the raw score ESs.

Hopefully, future research will help to establish how much difference in sampling (and thus variability between cases) that researchers can accommodate without creating notable problems for the synthesis of design-comparable ESs. In addition, we hope future research will provide options and guidance for how to handle consequential levels of heterogeneity. Perhaps, methods will be developed where the SD can be estimated using all cases across studies that were measured on a particular outcome variable, as opposed to estimating the variance within each primary study. Until additional research is conducted, we suggest that meta-analysts be aware that they are assuming that the effect sizes expressed on a given metric are exchangeable across studies, examine the studies with this in mind, and report findings with transparency about what is found.

### 2.4.2 ES Estimation: Homogeneity of Variance

The first assumption that comes to mind when thinking about synthesizing ESs is that the effect sizes are exchangeable in the SMD metric. However, that is not the only assumption required for conducting a synthesis of SCDs using design-comparable effect sizes. The estimation of design-comparable ESs assumes that the variation within a case is comparable across phases and comparable from case to case within a study. While often this assumption is reasonable, there are situations when it is not. For example, **?** utilized a multiple baseline (MB) design across three participants to examine the effects of math software with a game component on the off-task behaviors of students with ADHD. For the purpose of this methods guide, we extracted the data using Webplot Digitizer (**?**) and present it in Figure **??**. Results of our visual analysis suggest that the variance differs between the baseline and treatment phases, with treatment phases having less variation as the percentage of off-task behaviors decreased and approached zero. With count-based variables (e.g., raw counts or counts converted to percentages), variability often depends on the mean of the variable. Treatments that shift the mean tend to change the variance. In addition, we have noticed that substantial variability is common to unstructured baselines used in SCD studies. If the intervention phase provides more control (e.g., controlling for interventionist/peer attention, rate of reinforcement), we might expect some reduction in variance. Many single-case intervention studies are prone to this shift in variance across experimental phases. Until future research provides more concrete guidance about the best ways to proceed when uncovering heterogeneity, we recommend that meta-analysts be aware of what they are assuming and transparent about what is found.

In some circumstances, the outcomes chosen for SCD studies may be so unlike the outcomes used in group design research that synthesis across various study designs will not be feasible. In such case, we tend to see extreme violations of the homogeneity assumptions. For example, when all cases in a multiple-baseline study have baselines with consistent values of zero, we advise against trying to force the computation of a design-comparable ES. Only when the assumptions seem more reasonable and violations are more modest (or non-existent), should researchers entertain the use of design-comparable ES.

### 2.4.3 ES Estimation: Normality

Use of design-comparable ESs is also based on the assumption that the experimental observations are normally distributed around the case-specific trend lines. In some situations, the data may be consistent with this assumption of normality, whereas in other it will not. Examining the baseline phases for the math software intervention study by **?** in Figure **??**, observations appear distributed somewhat normally around the baseline means, with a pooled skewness value near zero ($sk = 0.15$) and a pooled kurtosis value near zero ($ku = -0.77$).

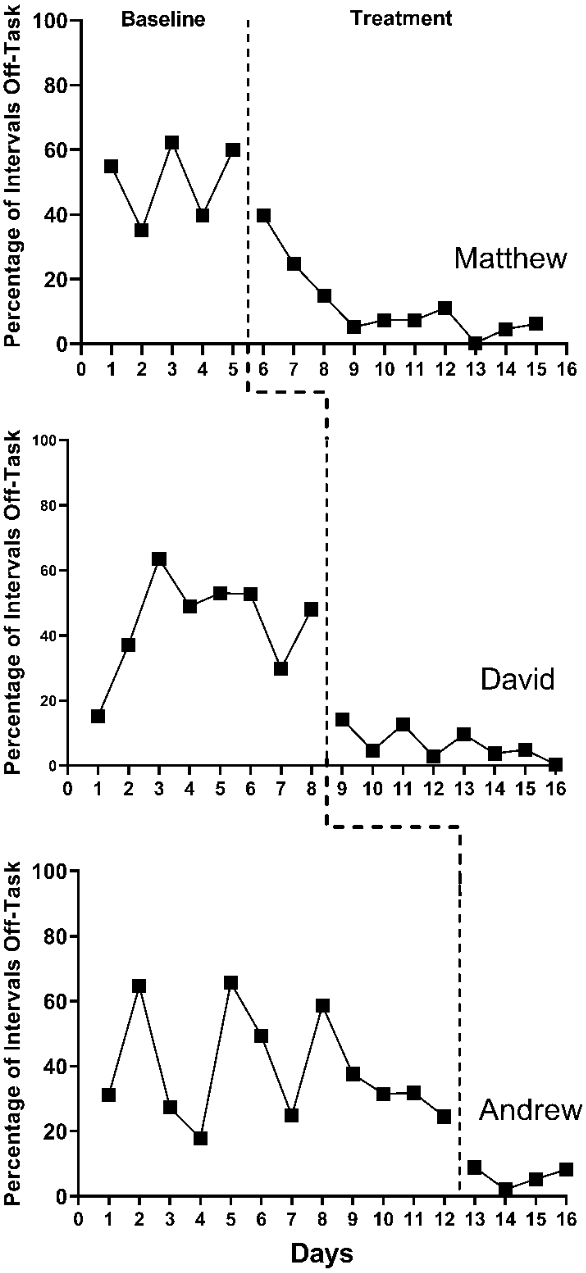

Figure 2.1: Multiple baseline design across three participants (Ota & DuPaul, 2002)

However, when we shift attention to the treatment phases, we see non-normality because the outcome is a percentage that has remained near the floor of 0% for much of the treatment phase ($sk = 1.86$; $ku = 4.99$).

With count-based variables, we often anticipate some departures from normality in addition to some differences in variance. These departures from what is assumed tend to be more pronounced when a count variable has a phase mean that is close to zero, or when a percentage variable has a phase mean close to 0% or close to 100%. However, when the counts are higher, or the percentage variable has a mean closer to 50%, the distributions tend to be more normal. We expect design-comparable ESs to tolerate some non-normality, but we need additional research to determine how much non-normality can be present before there are substantial consequences for the design-comparable ES.

We recommend that researchers consider normality, along with homogeneity, in their preliminary assessments of the primary studies to be included in a meta-analysis. As noted previously, there are some situations where the outcomes chosen for the single case studies may be so unlike the outcomes chosen for the group designs that synthesis across single case and group designs will not be feasible (e.g., SCDs where all cases in a multiple-baseline study have baselines with consistent values of zero). In such circumstances, we advise against trying to force the computation of a design-comparable ES. In other circumstances, design-comparable ESs can be estimated when the assumptions seem more reasonable and violations are more moderate or non-existent. Regardless of the findings, it is important to be transparent about the consistency of the data with the homogeneity and normality assumptions.

### 2.4.4 ES Estimation: Appropriate Structural Model

Finally, estimation of design-comparable ES requires making assumptions about the structural model for the data series collected in the SCD. These assumptions are ideally based on content expert knowledge of the intervention domain and dependent variables under review, as well as on visualization and calculation of descriptive statistics from the studies. We may assume that there is no trend in baseline, or there is a linear trend in baseline, or there is some form of nonlinear trend. We may assume that there is no trend in the treatment phase, or there is a linear or nonlinear trend in the treatment phase. Furthermore, we may assume that the parameters defining the baseline trajectory (e.g., level and slope) and the parameters defining the change in trajectory with treatment (e.g., the change in level and change in slope) are the same for all cases within a study, or that some of them are different for different cases. If we select a structural model that is inconsistent with our data, we expect biased design-comparable ES estimates.

As an example, we present the study of a writing intervention for post-secondary adults with intellectual and developmental disabilities that targeted the improvement of sentence construction (**?**). Figure **??** is a graphical depiction of the

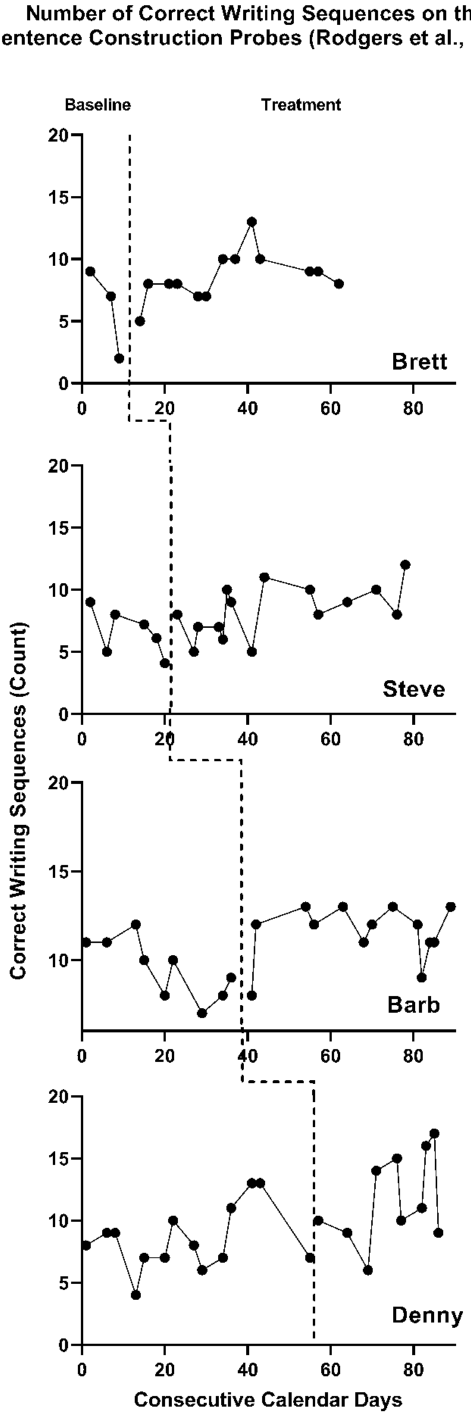

Figure 2.2: Multiple Baseline Across Participants (Rodgers et al., 2020)

outcome data and study design. Visual analysis of the baseline phase suggests potential baseline trends in accurate writing sequences. In our consideration of Denny, our visual analysis of correct writing sequences in baseline suggests the potential of an increasing trend in baseline, represented by the solid line in Figure **??**, which was estimated using ordinary least squares regression. In contrast, if we select a baseline model with no trends, the baseline projection is considerably different than the projected baseline when a linear trend is assumed (see the difference in the dotted lines through the treatment phase in Figure **??**). The effect estimate would be larger if we did not model the baseline trend in Figure **??**, because the observed treatment values are further above the projection based on no trend than the projection based on trend. Thus, whether or not we assume trends will have consequences for the ES estimates.

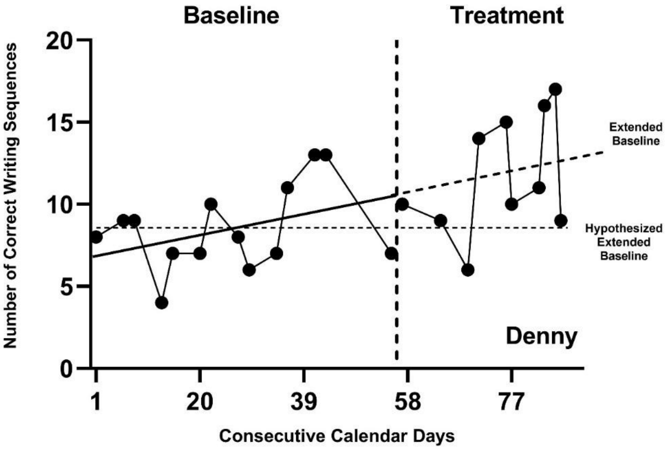

Figure 2.3: Hypothesized and Projected Baselines for Denny (Rodgers et al., 2020)

In thinking about an appropriate structural model, we recommend that meta-analysts start by carefully considering the area of research synthesis, including the outcome of interest, the participants in the studies, and what is expected in this area regarding baseline trends and treatment phase trends. In addition, we recommend that meta-analysts visually analyze the data from all primary studies to be included in the synthesis. In this visual analysis, the meta-analysts should assess the degree to which the data from the studies are reasonably consistent with expectations. If the data from the studies are consistent with the expectations about trends, these expectations can then be used to select among modeling options for design-comparable ESs. The next section details some of the commonly available modeling options for design-comparable ESs.

## 2.5 Modeling Options for Design-comparable Effect Size Estimation

We suggest that meta-analysts first consider the tenability of the homogeneity and normality assumptions. If violations are severe, researchers should reconsider if the outcomes from the single case studies are similar enough to the outcomes of the group studies to warrant synthesis using the design-comparable ES metric. If outcomes are substantially different across design types, it may be more reasonable to meta-analyze the SCD studies separately from the group studies and to use a different ES metric for the SCD studies. When violations are not too severe and outcomes appear more similar, then we suggest proceeding with the design-comparable ESs, while also noting what was found. If the decision is to proceed with design-comparable ESs, we suggest researchers next determine the predominant design used in the area of synthesis. Do most SCD studies in the research area tend to use reversal designs, such as ABAB designs, or do studies predominantly use MB across participants designs and/or multiple-probe (MP) across participant designs? When synthesizing reversal designs with design-comparable ESs, researchers are currently limited to models that assume there are no trends. For MB or MP designs, a variety of trend assumptions are feasible.

When the synthesis will include predominantly MB and/or MP designs, we suggest researchers clarify their expectations about trends given their understanding of the participants, context, and outcome being studied. We also recommend that the graphs of the data from the primary studies be analyzed visually for consistency with the trend expectations. Based on these considerations, we think it is helpful to determine which of the following sets of trend assumptions is most reasonable for the set of studies to be synthesized: a) no trends in baseline or treatment phases, b) no trends in baseline, but trends in treatment phases, or c) trends in baseline and different trends in treatment. After clarifying the trend assumptions, researchers next need to clarify assumptions about whether the treatment effect varies across the cases (Is the treatment effect expected to be the same for each case? Or is it expected that there will be differences in the response to intervention?) Again, we rely heavily on the expectations in the area of research synthesis, and visual analyses of the primary studies to determine if the data from those studies are reasonably consistent with the expectations. Figure **??** arranges these considerations into a series of decision rules that can be used select one of six common models for design-comparable ESs. Although there are other possibilities for specifying a model for design-comparable ESs, these six models cover a wide range of scenarios, can be estimated with the scdhlm software application (**?**), and include the models that have received the most attention in the methodological literature, as well as those that have been applied in meta-analyses of single case data.

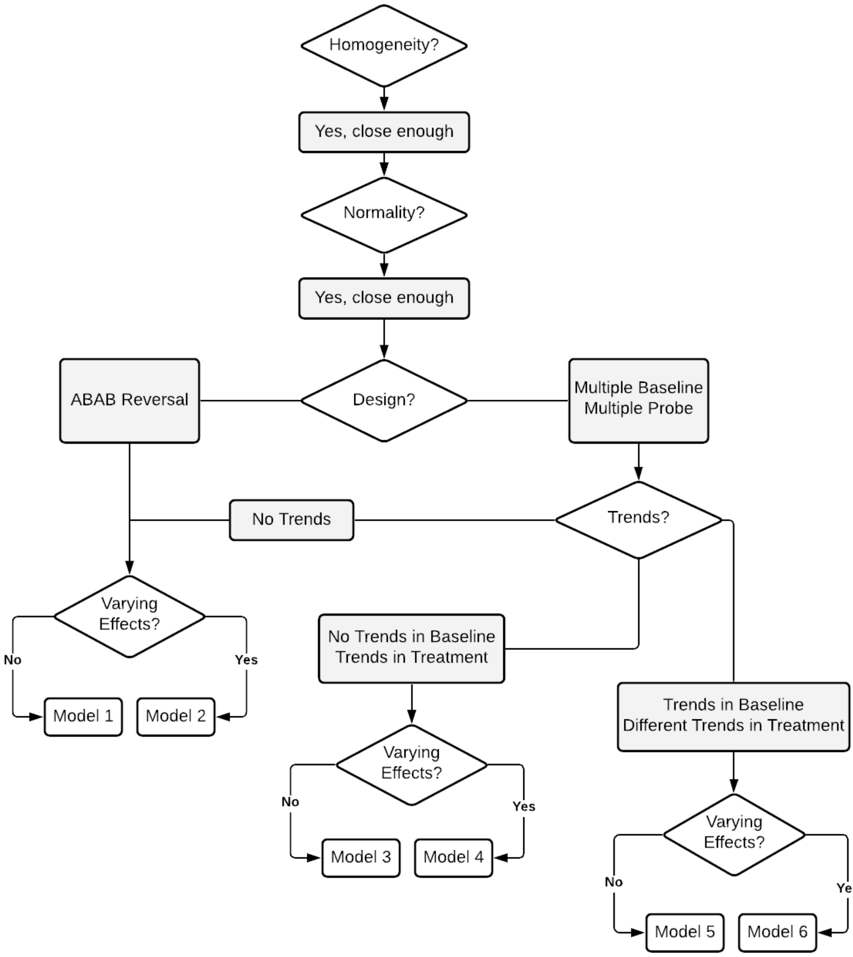

Figure 2.4: Flow Chart for the Selection of Design-comparable Effect Sizes